Structural Studies of MeCP2 in Complex with Methylated DNA

A Thesis Submitted for the Degree of Doctor of Philosophy

by

Kok Lian Ho, B.Sc. (Hons.), M.Sc.

Structural Biochemistry Group
The Institute of Structural and Molecular Biology
University of Edinburgh
October 2008

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DNA methylation is a common epigenetic mark that affects gene regulation, genomic stability and chromatin structure. In mammals, methylation is mainly found in the CpG dinucleotides. The CpG methylation signals can be recognised by the Methyl-CpG-Binding Protein (MBP) family which includes MeCP2, MBD1, MBD2, MBD3, MBD4 and Kiaso. MeCP2 and MBD1-4 (except mammalian MBD3) recognise methyl-CpG via their MBD domain whereas Kaiso interprets methylation through its Zn finger DNA binding domain. The TRD domains of MeCP2, MBD1 and MBD2 have been reported to recruit transcriptional co-repressors to the methylated DNA. A thymine DNA glycosylase domain is located at the C-terminal region of MBD4. This study concerns the molecular details of the methyl-CpG recognition by the MBD domain of MeCP2. To achieve this, the MeCP2 MBD domain has been expressed, purified and co-crystallised with a 20 bp DNA fragment from the *BDNF* promoter. The DNA-protein cocrystal diffracted X-rays to a maximum resolution of 2.5Å using synchrotron sources. It belongs to space group C2 with unit cell dimensions: a = 79.71Å, b = 53.60Å, c = 65.73Å, and $\beta = 132.1^{\circ}$. The X-ray structure of the MeCP2 MBD-DNA complex was solved using the SAD method. Structural analyses of the refined X-ray structure reveal that the methyl groups of the DNA make contact with a predominantly hydrophilic surface that includes tightly bound water molecules. From a structure of the MBD domain in MBD1, established by NMR, the binding specificity of the MBD domain had been thought to depend on hydrophobic interactions between the cytosine methyl groups and a hydrophobic patch within the MBD domain. The findings of this study suggest that MeCP2 recognises the hydration pattern of the major groove of methylated DNA rather than cytosine methylation *per se*. The X-ray structure also identifies a unique role of T158 and R106, the sites of the two most frequent Rett missense mutations. Both residues stabilise the tandem Asx-ST motif at the C-terminal region of MBD domain. Disruption of this tandem motif destabilises the DNA-protein interaction. The *BDNF* sequence in this study contains an AT run which displays unique properties of AT tract DNA. Previously, mutation of the AT run has been reported to decrease MeCP2 binding specificity. This study however demonstrated that a significant reduction can only be observed when both AT runs close to the methyl-CpG have been mutated. The X-ray structure of the MeCP2 MBD-DNA complex in this study rationalises the effects of the most common Rett mutations and provides a general model for methylated DNA binding that is dependent on structured water molecules.

DECLARATION

The work presented in this thesis is the original work of the author. This thesis has been composed by the author and has not been submitted in whole or in part for any other degree.

Kok Lian Ho

ACKNOWLEDGEMENTS

Firstly, I would like to express my gratitude to Prof. Malcolm Walkinshaw for giving me an opportunity to pursue a PhD in structural biochemistry in his laboratory. Thanks for his guidance particularly in X-ray crystallography and support throughout the course of this study. To my co-supervisor, Prof. Adrian Bird, I am most grateful for his supervisions and helpful discussion. I was delighted to share his in-depth knowledge in epigenetics particularly DNA methylation.

I would like to thank Dr. Iain McNae for his help in solving the crystal structure and his guidance in practical aspects of X-ray crystallography. Special thanks go to Dr. Julia Richardson for her help in data collection and helpful comments. I am thankful to Dr. Robert Klose and Dr. Lars Schmiedeberg for their help particularly in biophysical assays.

To many people on level 3 Swann Building; Sandra Bruce, Dr. Paul Taylor, Dr. Jacqueline Dornan, Dr. Hugh Morgan, Dr. Martin Wear, Dr. Matthew Nowicki, Connie Ludwig, Yi-gong Sheng, Kun-Yi Hsin, Dr. Nien-Jen Hu, Lu Zhou, Dr. Daphne Kan, Peter Brown, Jillian Adie, Elizabeth Blackburn and Ziad El-Hajj, thank for their helpful ideas and comments.

Thanks also go to Dr. Wen Siang Tan, Dr. Chyan Leong Ng, Dr. Wai Ling Kok and Swee Tin Ong for their encouragement and support.

To a special person in my life, Siew Yeong, thanks for her patience, understanding, support and love throughout these years.

Lastly, I am grateful to the Darwin Trust of Edinburgh for a postgraduate scholarship and the Wellcome Trust for funding this study.

献给我亲爱的父母。。。

Dedicated to my beloved parents…

ABBREVIATIONS

˚C	degree Celsius
Å	Ångström
BDNF	Brain-derived neurotrophic factor
bp	base pair
CCP4	Collaborative Computational Project Number 4
DNA	Deoxyribonucliec Acid
DNMT	DNA methyltransferase
DTT	Dithiothreitol
EDTA	ethylene diamine tetraacetic acid
EMSA	Electrophoretic mobility shift assay
FPLC	Fast Protein Liquid Chromatography
HDAC	Histone deacetylase
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HPLC	High-performance liquid chromatography
IMAC	Immobilized metal ion affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K	kelvin
kb	kilobase
K_d	The dissociation constant
kDa	KiloDalton
LB broth	Luria-Bertani broth
m5C	5' methyl cytosine
MAD	Multiwavelength anomalous dispersion
MBD	Methyl-CpG Binding Domain
MBD1	Methyl-CpG Binding Domain Protein 1
MBD2	Methyl-CpG Binding Domain Protein 2
MBD3	Methyl-CpG Binding Domain Protein 3
MBD4	Methyl-CpG Binding Domain Protein 4
MeCP2	Methyl-CpG Binding Domain Protein 2
MPB	Methyl-CpG Binding Protein
NDB	Nucleic Acid Database
NLS	Nuclear localisation signal
NTA	Nitrilotriacetic acid
OD	Optical density
PCM1	Protein Containing MBD 1
PCR	Polymerase chain reaction
PDB	Protein Database Bank
PEG	Polyethylene glycol
pI	Isoelectric point
RMSD	Root mean square deviation

RTT	Rett Syndrome
SAD	Single anomalous dispersion
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SeMet	seleno-methionine
TDG	Thymine DNA glycosylase
TRD	Transcriptional repression domain
V_m	Matthew's coefficient
$\gamma\ ^{32}$P dATP	gamma ^{32}P Deoxyadenosine triphosphate

CHAPTER 1. INTRODUCTION

1.1 DNA METHYLATION

DNA is made from four repeating building blocks; deoxyriboadenosine (dA), deoxyribocytosine (dC), deoxyriboguanosine (dG) and deoxyribothymidine (dT) (Figure 1-1). Their corresponding nucleosides are linked together by phosphodiester bonds to form polynucleotides. The DNA double helix consists of two strands of polynucleotides in which the nucleotide building blocks are paired in a uniquely complementary way (Watson and Crick, 1953). The human genome encodes over 25,000 proteins (Venter *et al.*, 2001). However, not all genes are equally expressed in different cells. Because distinct sets of genes are active in different cell types, protein expression profiles therefore must be regulated distinctively by cell specific chromatin structures which are in turn controlled by chromatin remodelling complexes. Chromatin remodelling complexes allow transcriptional regulators to gain access to their cognate DNA binding sites usually on the promoter region of genes. These specific DNA sequences are normally epigenetically marked so that they can be recognised by transcriptional regulators. Epigenetics is the study of heritable changes in gene function that occur without a change in DNA primary sequence (Bird, 2007). Epigenetic information is essential to determine distinct sets of cell type-specific gene activation (Brero *et al.*, 2006). At the chromatin level, epigenetic mechanisms include DNA methylation and modification of various functional groups of histones such as methylation, acetylation and phosphorylation [reviewed by Felsenfeld and Groudine (2003)].

DNA methylation is one of the most common forms of epigenetic signals that affect gene regulation, genomic stability and chromatin structure (Bird, 2002). DNA methylation occurs in many different organisms including prokaryotes, fungi, plants and animals. Addition of a methyl group at the 5' position of cytosine gives rise to 5'methyl cytosine (m5C) or at the 6' position of adenine; N6-methyl adenine (m6A). In prokaryotic systems, DNA methylation protects the host DNA from restriction enzyme digestion but not unmodified foreign DNA (Wilson and Murray, 1991). In eukaryotic cells, DNA methylation is implicated in more complex tasks including gene regulation, genomic imprinting, X inactivation, DNA replication and DNA

mismatch repair. The majority of the methylated DNA modifications are m5C, with some unicellular organisms showing low levels of m6A (Brero *et al.*, 2006). In mammals, methylation is mainly found at CpG dinucleotides. The human genome consists of approximately 1% of m5C, which results in 70-80% of CpG dinucleotides being methylated except CpG islands, which are present in approximately 60% of the promoter region of human genes (Bird, 2002). Therefore m5C is considered to be the 'fifth' nucleotide (Figure 1-1). Methylation in fungi and plants is not restricted to CpG dinucleotides but rather CpNpG sequences are the preferred methylation target [see review in (Selker, 1997; Tariq and Paszkowski, 2004)].

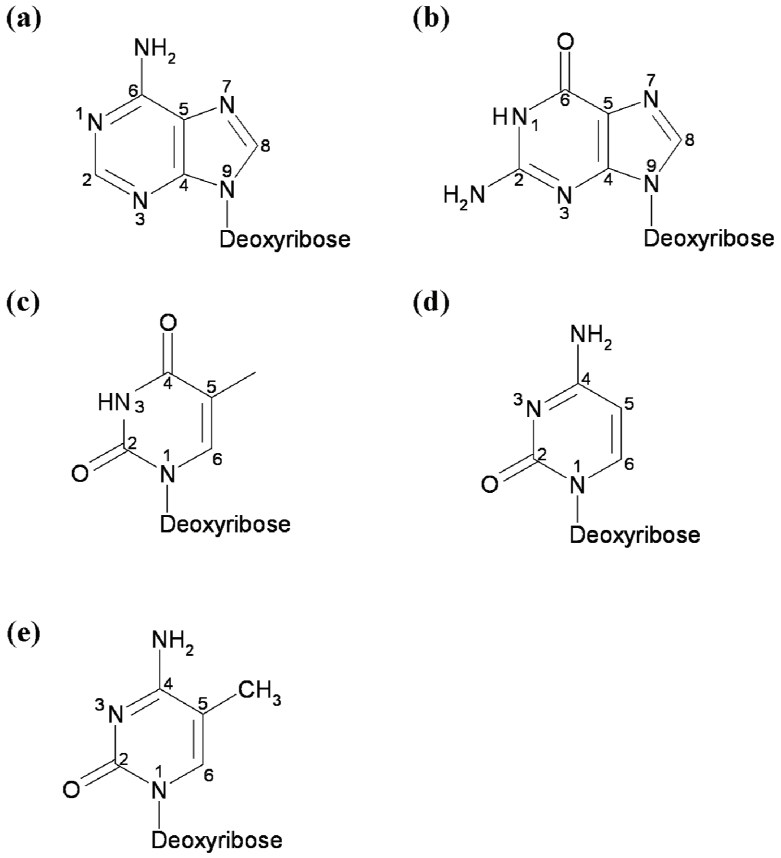


Figure 1-1 Building blocks of DNA
(a) deoxyadenosine (dA), (b) deoxyguanosine (dG), (c) deoxythymidine (dT), (d) deoxycytosine (dC) and (d) the fifth nucleotide; 5'methyl-deoxycytosine (m5C).

Methylation signals can be interpreted by Methyl-CpG Binding Protein (MBP) proteins which consist of MeCP2, MBD1, MBD2, MBD3, MBD4 and Kaiso (Klose and Bird, 2006). The classical methyl-CpG binding domain (MBD) containing proteins; MeCP2 and MBD1-4 (except mammalian MBD3); recognise symmetrically methylated 5'CpG3' (methyl-CpG) pairs via the MBD domain (Hendrich and Bird,

1998; Nan *et al.*, 1993) whereas Kaiso recognises the methylation signal through its zinc finger domain (Bird and Wolffe, 1999). Three MBD family members (MeCP2, MBD1 and MBD2) are able to recruit co-repressor complexes that can inhibit transcription with the aid of chromatin modifying enzymes. Mammalian MBD3 does not specifically bind methylated DNA and MBD4 functions primarily as a DNA G-T mismatch repair enzyme (Hendrich and Bird, 1998).

1.2 THE ROLES OF DNA METHYLATION

Silencing of genes and repetitive elements can be achieved by histone modification and chromatin structure alteration. These modifications are reversible and an additional mechanism is essential in order to silence these elements. DNA methylation is one of the stable modifications that can be inherited throughout cellular replication. The daughter cells therefore retain the same methylation patterns as their precursors which is necessary for silencing of repetitive elements in genome defence system, genomic imprinting, X-chromosome inactivation and cell development (Miranda and Jones, 2007).

Repetitive elements including retrotransposons and other parasitic elements which have accumulated in the mammalian genome during evolution represents at least 35% of the genome (Miranda and Jones, 2007; Yoder *et al.*, 1997b). Many of these repetitive elements contain a long terminal repeat promoter which allows the transcription of these sequences (Kochanek *et al.*, 1995; Yoder *et al.*, 1997b). Repetitive elements pose a potential threat to genomic stability and integrity as they can mediate recombination of non-allelic repeats, such as chromosomal rearrangements or translocations, and the active retrotransposons can integrate into genes and interfere with gene transcription and perhaps initiate retrotransposonal transcription (Robertson and Wolffe, 2000). Furthermore, examination of CpG methylation distribution within the genome has revealed that most of the parasitic DNA elements and retrotransposons, for instance; endogenous retroviruses, L1 elements and Alu elements; are considerably CpG enriched sequences (Colot and Rossignol, 1999; Yoder *et al.*, 1997b). Thus, DNA methylation has been proposed to be part of a genome defence system to silence expression of these repetitive elements and limit their spread in the genome (Yoder *et al.*, 1997b).

Genomic imprinting is an epigenetic mechanism of transcriptional regulation through expression of selected genes from one of the two parental alleles (Delaval and Feil, 2004). The requirement for imprinting in mammals was first discovered following nuclear transplantation experiments. Mouse embryos that contained only the maternal (parthenogenoes) or paternal genomes (androgenotes) were retarded in development (McGrath and Solter, 1984; Surani *et al.*, 1984). These experiments highlighted the importance of complementarity of male and female genomes in order to achieve normal development. Most of the imprinted genes are clustered in the genome rather than lone genes. To date, about 80 imprinted genes have been identified in human and mouse (http://www.mgu.har.mrc.ac.uk/research/imprinting/index.html).

Expression of imprinted genes within a cluster is controlled by a CpG rich sequence element known as the imprinting control region (ICR) which is usually composed of several kilobases and is regulated by DNA methylation (Feil and Berger, 2007). DNA methylation of most ICRs is established during oogenesis (female gametogenesis) while only some of them are established during spermatogenesis (male gametogenesis). Following fertilisation, these methylation marks are maintained throughout development and govern the allelic expression of the imprinted genes. One of the most well studied imprinted gene clusters is insulin-like growth factor-2 gene (*Igf*2) and H19 gene on distal mouse chromosome 7 (Bartolomei *et al.*, 1991; DeChiara *et al.*, 1991). The ICR of *Igf*2-*H19* cluster is paternally methylated while the unmethylated maternal copy is bound by an insulator factor namely CCCTC binding factor (CTCF). An insulator is a DNA sequence located between promoter and enhancer which blocks the interaction between the two elements (Bell *et al.*, 1999). The binding of CTCF is eliminated by CpG methylation within the CTCF binding site *in vitro* (Bell and Felsenfeld, 2000). Figure 1-2 shows the reciprocal imprinting of *Igf*2-*H19* genes, CTCF binds to the CpG rich non-methylated ICR on the maternal allele and insulates the interaction between *Igf*2 promoter and the enhancers located at the 3' region of *H19* and consequently, allows the *H19* promoter to exclusively access the enhancers (Verona *et al.*, 2003). The ICR methylation on paternal allele blocks CTCF binding, therefore preventing the formation of insulator and this architecture allows *Igf*2 promoter access to the downstream enhancers which are required for *Igf*2 expression. Together with the hypermethylation of *H19* promoter on the paternal

allele, the methylated ICR represses the transcription of *H19*. MeCP2 (Drewell *et al.*, 2002) and, recently, MBD3 (Reese *et al.*, 2007) have been demonstrated to be involved in silencing at the paternal H19 allele.

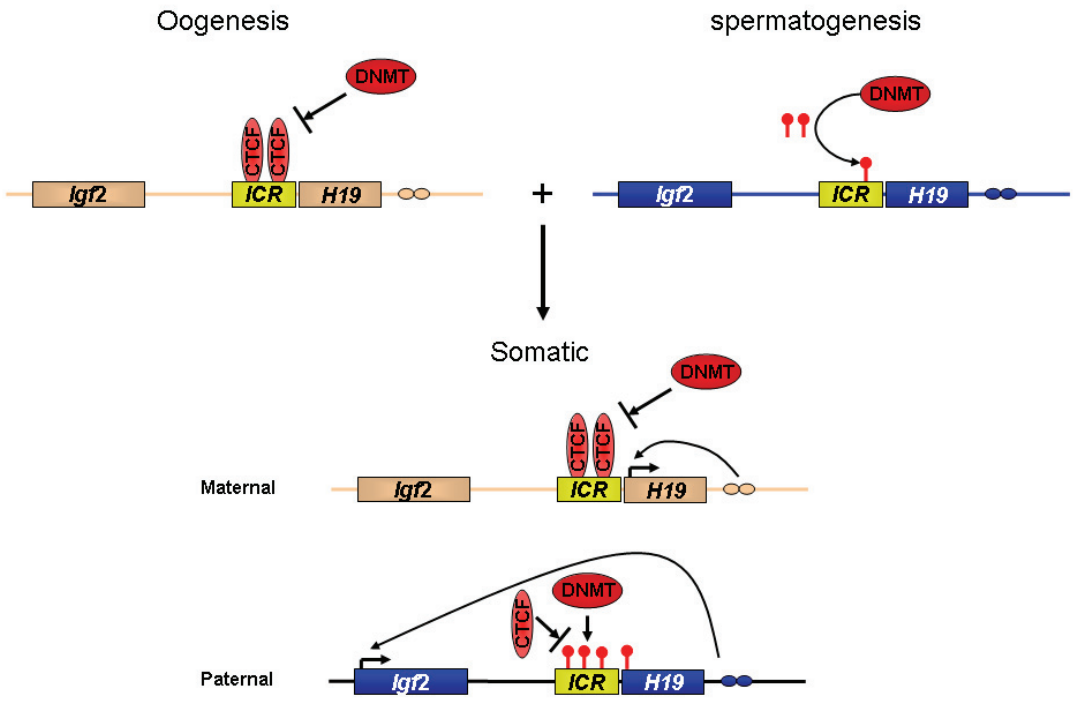


Figure 1-2 Imprinting in the *Igf2-H19* cluster
The male ICR of *Igf2-H19* cluster acquires methylation during the late stage of spermatogenesis. By contrast, the female ICR is protected from methylation by the bound zinc finger protein CTCF. After fertilisation, this epigenetic mark is maintained throughout the cell development. The binding of CTCF on the chromosome creates a boundary which prevents the interaction between *Igf2* promoter and the enhancers. Accessibility of the *H19* gene to the downstream enhancer allows the transcription of the *H19* gene. On the other hand, ICR methylation eliminates the binding of CTCF and therefore allows the *Igf2* gene to access the enhancer. As a result, the *Igf2* gene is expressed paternally. Paternal *H19* gene is repressed because of hypermethylation spreading from the upstream methylated ICR (diagram was adapted from Delaval and Feil, 2004).

In mammals, dosage compensation of gene expression between males (XY) and females (XX) is established by random inactivation of one of the two X-chromosomes in females. This process is known as X chromosome inactivation (XCI) (Lyon, 1961). Mammalian XCI is regulated by reciprocal expression of 17kb non-coding *Xist* (X-inactive specific transcript) and its antisense counterpart *Tsix* RNAs, both encoded by a region called X inactivation centre (*Xic*) on the X chromosome (Brockdorff *et al.*, 1992; Brown *et al.*, 1992; Lee and Lu, 1999). XCI is initiated by expression of

non-coding *Xist* gene from the inactive X chromosome (Xi). The accumulation of *Xist* RNA transcripts silence the chromosome in *Cis* (Marahrens *et al.*, 1997; Penny *et al.*, 1996). The *Xist* accumulation, however, is not required for continuing maintenance of the Xi silencing (Wutz and Jaenisch, 2000) but other factors such as chromatin modulators are involved. On active chromosome (Xa), up-regulation of *Xist* RNA is inhibited by the expression of the non-coding *Tsix* RNA (Lee and Lu, 1999; Stavropoulos *et al.*, 2001). In regulating *Tsix* expression on Xa, an insulator CTCF binds near to the promoter of *Tsix* (Figure 1-3), which enables the expression of *Tsix* gene due to its accessibility to the enhancers (Chao *et al.*, 2002) and insulates the *Xist* promoter from accessing the enhancers. On the Xi, DNA methylation on the CTCF array prevents the binding of CTCF on *Tsix* and thus allows *Xist* gene expression.

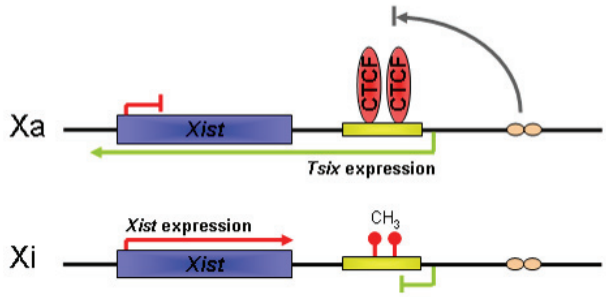


Figure 1-3 X chromosome inactivation
On Xa, CTCF binds to the unmethylated region near the *Tsix* promoter, in which, *Xist* promoter is insulated from accessing the downstream enhancers. Conversely, the interaction of *Tsix* promoter and enhancers allows *Tsix* transcription, which possibly induces repressive chromatin modification across the *Xist* promoter region and represses *Xist* expression. On Xi, CTCF binding is prevented by methylated CpGs near the *Tsix* promoter. This configuration enables the *Xist* promoter to access the downstream enhancers and create a repressive state for *Tsix* expression (diagram adapted from Chao et al., 2002).

Recent evidence suggests that *Xist* is regulated by *Tsix* expression through modification of the chromatin structure in the *Xist* promoter region (Ohhata *et al.*, 2008; Sado *et al.*, 2005). It is known that the promoter region of active *Xist* gene on Xi is hypomethylated, whereas that of the transcriptionally inactive *Xist* gene on Xa is hypermethylated (Norris *et al.*, 1994). Loss of *Tsix* transcription across the *Xist* promoter reduces CpG methylation and also causes histone modification in the 5' *Xist* region (Ohhata *et al.*, 2008; Sado *et al.*, 2005), indicating that a modified chromatin structure is required to promote DNA methylation.

1.3 ESTABLISHMENT AND MAINTENANCE OF DNA METHYLATION

To date, five mammalian DNA cytosine methyltransferases (DNMTs) have been discovered, which can be divided into three families; DNMT1, DNMT2 and DNMT3 (Figure 1-4) (Bestor, 2000). Based on their preferred DNA substrate, DNMTs can generally be grouped into two classes; *de novo* and maintenance DNMTs. The *de novo* DNA methyltransferases DNMT3A and DNMT3B are mainly responsible for introducing cytosine methylation at previously unmethylated CpG sites, whereas the maintenance methyltransferase DNMT1 duplicates the DNA methylation pattern onto the new DNA strand during DNA replication. The fourth DNA methyltransferase (DNMT2) shows little DNA methyltransferase activity *in vitro* even though it is structurally similar to prokaryotic DNA methyltransferase (Hermann *et al.*, 2003). Targeted deletion of DNMT2 does not affect DNA methylation globally in embryonic stem cells, indicating that this enzyme is not essential for *de novo* or maintenance of DNA methylation (Okano *et al.*, 1998b). DNMT3L is a paralogue of DNMT3A and DNMT3B, which enhances the *de novo* DNMTs catalytic activities but does not physically associate with methylated DNA (Suetake *et al.*, 2004). Nevertheless, the crystal structure of bacterial DNA cytosine methyltransferase complexed with substrate DNA suggests a similar enzymatic mechanism for mammalian DNMTs.

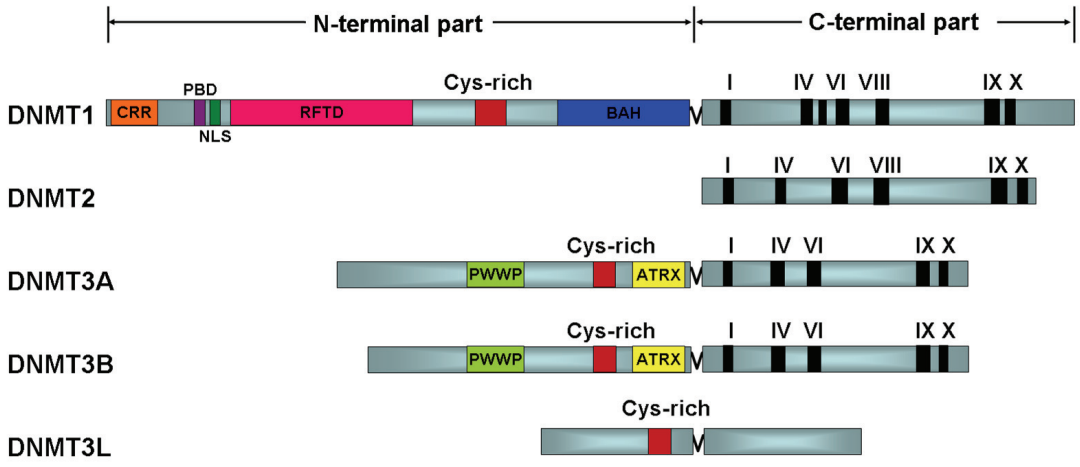


Figure 1-4: Schematic representative of mammalian DNA methylfransferases
The N-terminal part of DNMT1 consists of regulatory elements: charge rich region (CRR), PCNA binding domain (PBD); nuclear localisation signal (NLS), replication foci targeting domain (RFTD), CxxC motifs (Cys-rich), bromo-adjacent homology domains (BAH). In addition to Cys-rich domain, the N-terminal part of DNMT3A/B also contains PWWP domain for protein-protein interaction and ATRX domain (involved in HDAC interaction). The N- and C- terminal of DNMTs are linked by Gly-Lys dipeptides. The C-terminal catalytic region is related to bacterial DNA methyltransferases. Black bars indicate conserved catalytic motifs among DNMT families (diagram was adapted from Bestor, 2000).

1.3.1 Mechanism of DNA methylation

The cocrystal of M.*Hhal* (Klimasauskas *et al.*, 1994) (Figure 1-5) and M.*Hae*III (Reinisch *et al.*, 1995) methyltransferases complexed with DNA revealed the catalytic mechanism of DNA methyltransferase. Both crystal structures display high similarity with a C_α RMSD of 1.18Å over 119 amino acids. As shown in Figure 1-6, the prokaryotic DNA cytosine methyltransferases catalyse the transfer of methyl group from S-adenosyl-L-methionine (SAM) to the C5 position of the cytosine. In order to capture the intermediate catalytic state, the hydrogen atom at C5 was replaced by a fluorine atom which cannot be released as free F^+. The target cytosine is flipped out of the DNA helix and placed in the active site of the enzyme where the protein is covalently linked to the DNA. The sulphur atom of nucleophile Cys81 forms a covalent thioester bond to C6 of cytosine (Figure 1-5). The methyl group of SAM is then transferred to the C5 position of cytosine and the cofactor is converted to S-adenosyl-L-homocysteine. Abstraction of the C5 proton allows reformation of a carbon-carbon double bond. The methylation is then terminated by β-elimination. The pyrimidine ring of cytosine is held in the catalytic pocket by a number of conserved amino acids among the DNMT families. This suggests a similar catalytic mechanism for mammalian cytosine methyltransferases.

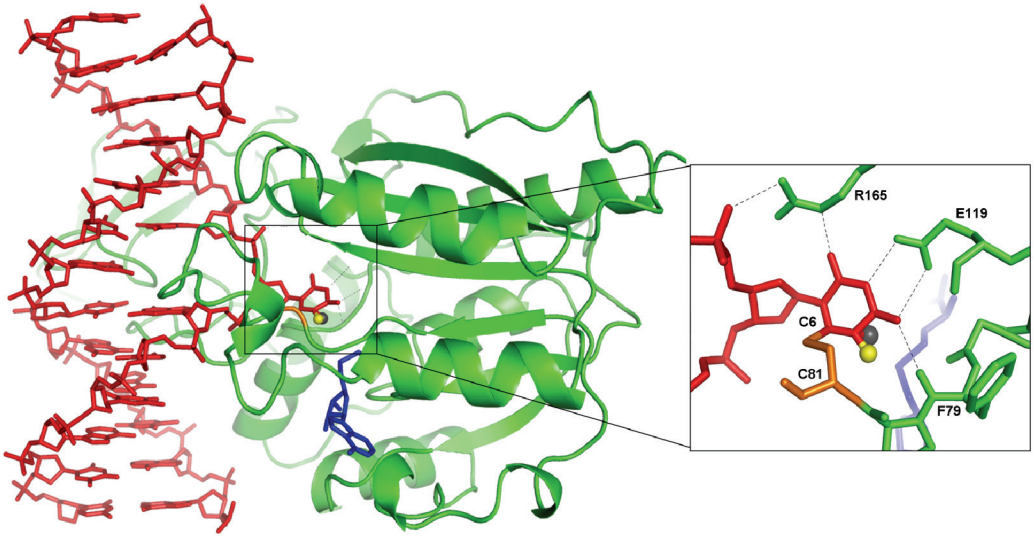


Figure 1-5 X-ray structure of the bacterial M. *Hha*I complexed with DNA
The target cytosine is held in the catalytic pocket by side-chains of R165, E119 and main-chain NH of F79. C81 (orange) forms a covalent thioester bond with C6 of cytosine (inset). The methyl group (grey sphere) at the C5 position of cytosine is transferred from SAM and the cofactor is then converted to S-adenosyl-L-homocysteine (blue). The irreplaceable fluorine atom (yellow sphere) halts the catalytic reaction (diagram adapted from Klimasauskas *et al.*, 1994).

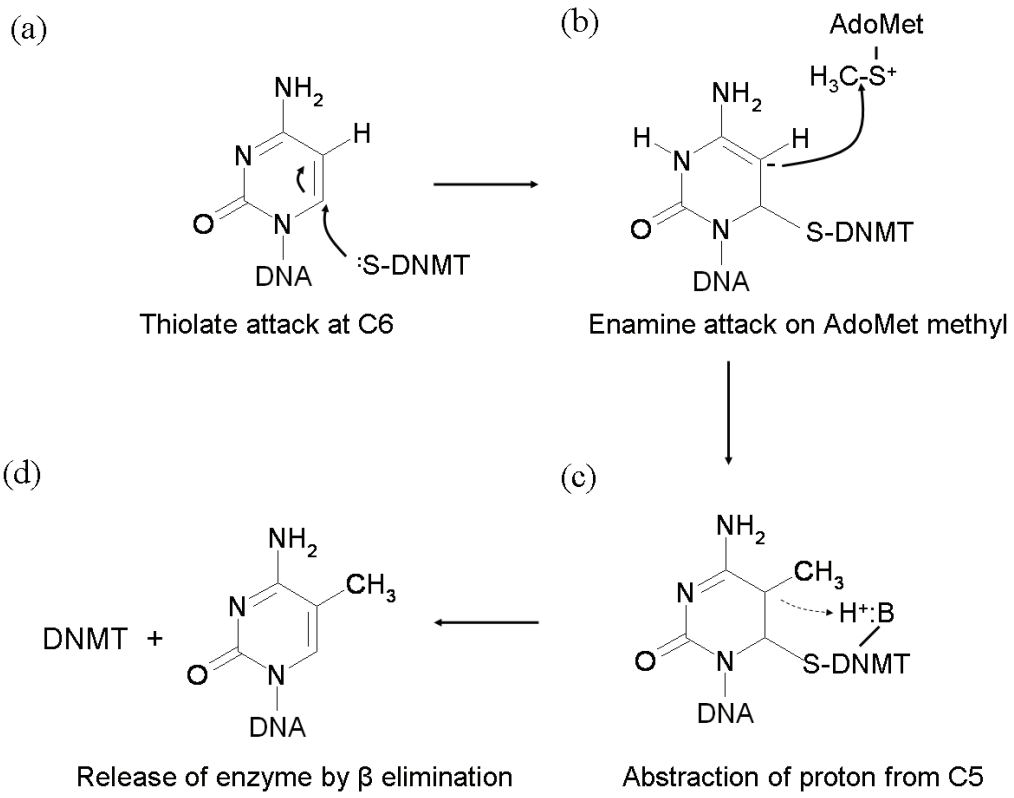


Figure 1-6 Schematic representation of DNMT catalytic pathway
(a) The catalytic mechanism involves nucleophilic attack on C6 of cytosine by a conserved cystein residue to generate a covalently linked intermediate. (b) SAM is converted to S-adenosyl-L-homocysteine by transferring the methyl group to C5 of cytosine. (c) Abstraction of C5 proton allows reformation of carbon-carbon double bond. (d) methyl-cytosine is released from the DNMT by β-elimination (diagram adapted from Bestor, 2000).

1.3.2 DNMT1

DNMT1 is the first mammalian DNA cytosine methyltransferase to be cloned and biophysically assayed (Bestor *et al.*, 1988). It has a 5 to 30-fold preference for hemimethylated DNA (Yoder *et al.*, 1997a), therefore, it has been suggested to play a role in maintaining methylation patterns. Yoder et al. (1997a) also reported that the specificity of DNMT1 is confined to the CpG motif rather than the primary DNA sequence or density of CpG dinucleotides. This protein contains 1620 amino acids and can be divided into two major domains; a large N-terminal domain with regulatory function and a small C-terminal domain carrying catalytic function (Hermann *et al.*, 2004) (Figure 1-4). The regulatory N-terminal domain bears different motifs, like the charge-rich domain that interacts with the Dmap1 transcriptional repressor (Rountree *et al.*, 2000), a nuclear localisation signal (NLS), a proliferating cell nuclear antigen (PCNA) binding motif, a Cys rich (CxxCxxC) Zn^{2+} binding

domain, a polybromo-1 region containing two BAH motifs suggested to play a role in protein-protein interaction, and a DNA replication foci targeting sequence that is responsible for DNA replication in eukaryotic cells (Hermann *et al.*, 2004). In addition, DNMT1 also directly interacts with histone modifying enzymes such as HDAC1 and HDAC2, associating with MBD2, MBD3 (Tatematsu *et al.*, 2000) and MeCP2 (Kimura and Shiota, 2003), the heterochromatin binding protein HP1 and SUV39H1 (Fuks *et al.*, 2003a). All these interactions are required for transcriptional repression.

A Lys and Gly rich sequence connects the N- and C-terminal regions of DNMT1. The C-terminal region contains the catalytic domain of DNMT1, which is closely related to prokaryotic DNA cytosine methyltranferases rather than mammalian DNA methyltransferases of DNMT2 and DNMT3 (Bestor, 2000). The C-terminal region alone however is catalytically inactive and a large part of the N-terminal domain is required to retain its methylating activity (Margot *et al.*, 2000; Zimmermann *et al.*, 1997). This region contains all conserved motifs where a common catalytic architecture is shared with other DNMT families. The catalytic mechanism of all DNMTs that modify the 5' position of the pyrimidine ring appear to use a similar mechanism to that described in Figure 1-6. Furthermore, most of the enzymes have a conserved prolylcysteinyl active site that provides the cysteine thiolate for the cytosine methylation mechanism (Figure 1-6) (Bestor and Verdine, 1994). Comparison of DNA cytosine methyltransfases revealed that the enzymes carry characteristic sequence motifs (Figure 1-4), namely motifs I to X, six of which are highly conserved (Lauster *et al.*, 1989; Posfai *et al.*, 1989). Motifs I and X form the Ado-Met binding site, motif IV contains the Pro-Cys dipeptide that provides the thiolate at C6 of the pyrimidine ring, motif VI bearing a glutamate that protonates the cytosine C3 position, and motif IX plays a structural role in maintaining the DNA-protein recognition surface in the DNA major groove (Bestor and Verdine, 1994).

1.3.3 DNMT2

DNMT2 was discovered a decade after DNMT1 was identified and proved to be an enigmatic enzyme. This enzyme is a relatively small protein of 391 amino acids and lacks the large N-terminus present in the DNMT1 and DNMT3 families (Figure 1-4).

The first mammalian DNMT2 identified using a cDNA database search (Yoder and Bestor, 1998) shares sequence similarity to pmt1$^+$ of *Schizosaccharomyces pombe*, an organism that is believed not to methylate its DNA (Wilkinson *et al.*, 1995). A DNMT2 knockout in mouse is a homologue of pmt1$^+$ and had no obvious effect on genomic methylation patterns in embryonic stem cells or on newly integrated retroviral DNA (Okano *et al.*, 1998b), indicating that DNMT2 is not required for *de novo* methylation or maintenance of methylation patterns. Recently, DNMT2 has been reported to have a novel tRNA methyltransferase activity at the tRNA^{Asp} anticodon (Goll *et al.*, 2006). Purified DNMT2 from human and *Drosophila melanogaster* can methylate cytosine in tRNA^{Asp} but not in DNA (Goll *et al.*, 2006), indicating that the DNMT2 may function as an RNA methyltransferase despite the fact that its catalytic motifs are highly conserved with other DNMT families. The crystal structure of human DNMT2 contains the catalytic sequence motif and shows a high similarity to M.*Hha*I but methyltransferase activity was not detected (Dong *et al.*, 2001).

1.3.4 DNMT3

Using a cDNA database search, mammalian DNMT3A and DNMT3B were identified, which show little sequence similarity to mammalian DNMT1 and DNMT2 (Okano *et al.*, 1998a). Nevertheless, the general architecture of DNMT3 resembles DNMT1 with the C-terminal region containing all the MTase motifs (Figure 1-4). DNMT3A and DNMT3B are closely related proteins that bear a PWWP domain and a CxxCxxC domain at the N-terminal tail related to that of DNMT1, ATRX and MBD1 (Xie *et al.*, 1999). Mammalian DNMT3A and DNMT3B are highly expressed during early development of mammalian embryos. Knockout of DNMT3A and DNMT3B in mouse ES cell abolish *de novo* methylation but has no effect on maintenance of imprinting methylation patterns (Okano *et al.*, 1999), indicating that both enzymes are required for mouse *de novo* methylation and development. However, DNMT3A and DNMT3B have also been suggested to perform maintenance methylation as these enzymes compensate for insufficient methyltransferase activity by DNMT1 (Liang *et al.*, 2002). Moreover, DNMT1 and DNMT3B knockouts, individually, in human cancer cells exhibited a slight reduction of overall genomic methylation, whereas double knockout of both DNMT1 and DNMT3B nearly eliminated methyltransferase activities including demethylation of repeated sequences, loss of *IGF2* imprinting,

abrogation of silencing of tumour suppressor gene, $p16^{INK4a}$ (Rhee *et al.*, 2002; Rhee *et al.*, 2000). These results demonstrated that a co-operative role of DNMT1 and DNMT3B is established to accomplish genomic methylation. *In vitro* studies showed that the methylation target of DNMT3A and DNMT3B is not restricted to cytosine at CpG sites, but non-CpG sites are also methylated by both enzymes (Gowher and Jeltsch, 2001; Ramsahoye *et al.*, 2000; Suetake *et al.*, 2003).

DNMT3L (DNA methyltransferase 3-like) is the third homologue of the DNMT3 family that is expressed specifically during gametogenesis in germ cells (Aapola *et al.*, 2000). DNMT3L is related to DNMT3A and DNMT3B in both C- and N-terminals with a conserved CxxCxxC domain but lack of a PWWD domain. However, the key residues within the catalytic motifs are not conserved (Figure 1-4) and thus the protein has not been shown to contain methylation activity. DNMT3L is essential for establishment of a subset of methylation patterns for genomic imprinting in male and female germ cells (Bourc'his *et al.*, 2001). Therefore, DNMT3L has been suggested to play a role in modulating methylation activities of DNMT3A and DNMT3B in establishment of maternal imprints (Bourc'his *et al.*, 2001). DNMT3L increases the DNA methylation activity of DNMT3A and DNMT3B but not DNMT1 (Suetake *et al.*, 2004). Inactivating both DNMT3A and DNMT3B abolishes *de novo* methylation in mouse embryos (Okano *et al.*, 1999). Conditional knockout of DNMT3A in mouse embryos showed lack of methylation at imprinted loci while conditional knockout of DNMT3B showed no apparent phenotypic changes compared with the wildtype animal (Kaneda *et al.*, 2004). Correspondingly, the DNMT3A knockout was indistinguishable from the DNMT3L knockout animal, thus indicating that DNMT3A and DNMT3L are both required for the methylation of most imprinted foci in germ cells (Kaneda *et al.*, 2004). Crystallographic studies of DNMT3L with DNMT3A2 (an alternative splice variant of DNMT3A), revealed that the CxxCxxC domain of DNMT3L connects ummethylated Lys4 of histone H3 tail to *de novo* methylation of DNA (Ooi *et al.*, 2007). This result indicates that DNMT3L recognises histone H3 tails that are unmethylated at Lys4 and induces *de novo* methylation by recruitment or activation of DNMT3A2.

1.4 INTERPRETATION OF DNA METHYLATION

1.4.1 Methyl-CpG binding activities

The first evidence indicating that methyl-CpG binding proteins bind to methylated DNA came from restriction endonuclease accessibility assays in isolated cellular nuclei (Antequera *et al.*, 1989). Protein bound nuclei were protected from restriction enzyme digestion at methyl-CpG sites, but naked DNA was easily digested. MeCP1 was the first methyl-CpG binding activity identified that complexed with methyl-CpG moieties (at least 15 methylated CpG) regardless of the primary DNA sequence (Meehan *et al.*, 1989). Four years later, a second methyl-CpG binding activity was reported, namely MeCP2 (Lewis *et al.*, 1992), which was abundant in brain extract. MeCP2, however, is the first methyl-CpG binding protein to be cloned and characterised (Lewis *et al.*, 1992).

1.4.2 Identification of methyl-CpG binding protein family

The Methyl-CpG binding domain (MBD) was initially mapped to an 85 amino acid polypeptide of MeCP2 that binds exclusively to symmetrical methyl-CpG using southwestern blot analysis (Nan *et al.*, 1993). The MBD sequence of MeCP2 was then used as a search template in cDNA databases which subsequently identified a second methyl binding protein, namely Protein Containing MBD 1 (PCM1) (Cross *et al.*, 1997). Recombinant PCM1 bound specifically to methyl-CpG and later was renamed to MBD1 (Hendrich and Bird, 1998). Using a similar approach, mammalian MBD2, MBD3 and MBD4 were subsequently identified (Hendrich and Bird, 1998). Mammalian MBD2 and MBD4 bind specifically to methyl-CpG but mammalian MBD3 does not recognise methylated DNA *in vitro* and *in vivo* (Hendrich and Bird, 1998). Based on their primary putative MBD sequences, the MBDs can be divided into two subgroups. The MBD sequence of MeCP2 shows a high similarity with MBD4 while MBD1, MBD2 and MBD3 are more similar to one another (Hendrich and Bird, 1998).

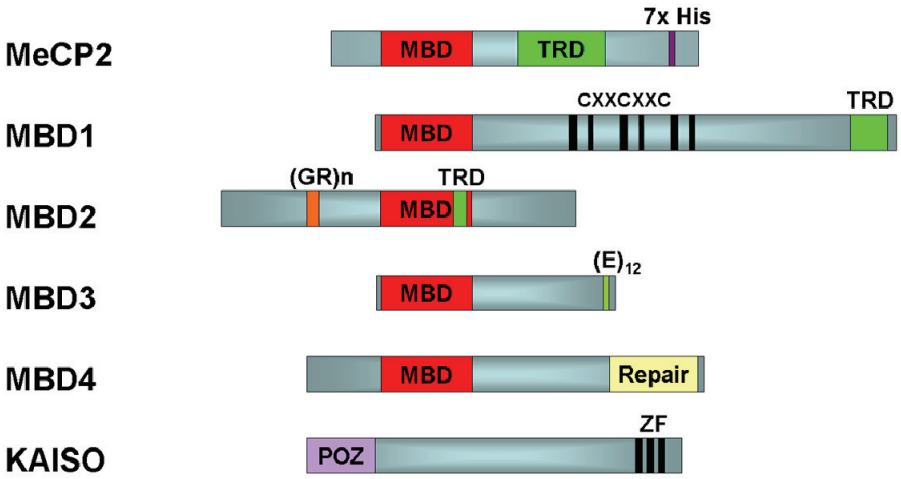


Figure 1-7: The methyl-CpG binding protein family
Classical methyl-CpG binding domain proteins; MeCP2, MBD1, MBD2, MBD3 and MBD4 interacting DNA through their MBD domain while Kaiso binds methyl-CpG via its ZF domain. MBD1 contains 3 CxxC zinc binding domain, one of which bind specifically to non-methylated CpG dinucleotides and a C-terminal TRD domain. MBD2 bearing a GR repeat RNA binding domain and a TRD domain which overlapped the MBD domain. MBD3 contains a polyglutamate and a well conserved MBD. MBD4 binds to methylated DNA via its MBD domain and the C-terminal region carrying a GT mismatch repair glycosylase.

1.4.3 MBD1

MBD1 was originally identified as a component of the transcriptional repressor MeCP1 (Cross *et al.*, 1997). MBD1 is characterised by the presence of 3 zinc coordinating cysteine-rich (CxxC) domains (Cross *et al.*, 1997; Fujita *et al.*, 1999) (Figure 1-7). A similar motif is also found in DNMT1 and mammalian HRX protein (Cross *et al.*, 1997). To date, five isoforms have been identified in human which are alternatively spliced in the region of the third CxxC (CxxC3) domain and the C-terminus of MBD1 (Fujita *et al.*, 1999; Jorgensen *et al.*, 2004). The MBD domain located at the N-terminus of MBD1 and a short TRD domain comprises 32 amino acids located at the C-terminal region (Fujita *et al.*, 1999; Ng *et al.*, 2000).

The MBD domain of MBD1 specifically binds methyl-CpG but the CxxC domains has been shown to interact with non-methylated DNA (Jorgensen *et al.*, 2004). The major isoform of mouse cells, MBD1a, containing a CxxC3 domain represses non-methylated reporter genes by targeting non-methylated CpG sites, for which, an intact MBD domain is dispensable (Jorgensen *et al.*, 2004). A similar non-methylated CpG binding has been observed in the CpG-binding protein (CGBP) and Mixed Lineage

Leukemia (MLL), which discriminates against DNA methylation (Birke *et al.*, 2002; Lee *et al.*, 2001).

The TRD domain of MBD1 is able to repress transcription at a distance up to 2 kb upstream of a promoter in a histone deacetylase independent manner (Ng *et al.*, 2000). MBD1 recruits histone H3-K9 methylase SETDB1 to a large subunit of Chromatin Assembly Factor (CAF-1) to form an S phase-specific CAF-1/MBD1/SETDB1 complex that facilitates methylation at H3-K9 during replication-coupled chromatin assembly (Sarraf and Stancheva, 2004). The histone H3 methylase (Suv39h1) and the methyl lysine-binding protein (HP1) heterochromatin complex interact with the MBD domain of MBD1 but not the TRD domain (Fujita *et al.*, 2003b). MBD1 associates with histone deacetylases through Suv39h1 (Fujita *et al.*, 2003b). MBD1-containing chromatin-associated factor 1 (MCAF1), also known as human homologue of murine ATFa-associated modulator (AM) interacts with the TRD domain (Fujita *et al.*, 2003a; Ichimura *et al.*, 2005). Therefore, MBD1 represses transcription by recruiting HP1, Suv39h1 and histone deacetylases via the MBD (Fujita *et al.*, 2003b) and MCAF1 via the TRD domain (Fujita *et al.*, 2003a). Together with SETDB1, MCAF enhances transcriptional repression by MBD1 (Ichimura *et al.*, 2005). MBD1 is also associated with the DNA damage protein methylpurine-DNA glycosylase in DNA repair, suggesting a direct DNA damage sensing role by MBD1 (Watanabe *et al.*, 2003).

1.4.4 MBD2 and MBD3

There are two major isoforms of MBD2; the full length MBD2a and a truncated form; MBD2b; the latter lacking the N-terminal 140 amino acids (Hendrich and Bird, 1998). The N-terminal region of MBD2a contains a GR_n repeat which has been reported to interact with RNA (Jeffery and Nakielny, 2004). MBD2b binds DNA *in vitro* in a methylation-dependent manner (Hendrich and Bird, 1998). Mapping of the MBD2b portion required for transcriptional repression revealed that the TRD domain partially overlapped the MBD domain (Boeke *et al.*, 2000) (Figure 1-7). MBD2 in HeLa cells associates with histone deacetylase (HDAC) in the MeCP1 complex and the transcriptional repression is relieved by the deactylase inhibitor trichostatin (TSA) (Ng *et al.*, 1999). Being a transcriptional repressor, MBD2b has also been reported to

contain specific demethylase activity for methylated CpG (Bhattacharya *et al.*, 1999), although this has been questioned in two independent reports (Ng *et al.*, 1999; Wade *et al.*, 1999).

MBD2 can recognise a single methylated CpG pair but it has been reported distinct from MeCP2 in that it prefers a densely methylated DNA (at least 12 mCpG pairs) for productive binding (Lewis *et al.*, 1992; Meehan *et al.*, 1989). MBD2 acts as a methyl-binding domain in the MeCP1 complex which also contains histone deacethylases (HDAC1/2) and RbAp46/48 proteins (Ng *et al.*, 1999), allowing MBD2 to target HDACs/chromatin remodelling activities at methylated loci (Feng and Zhang, 2001). MBD2 also recruits MBD3 containing Mi-2/NuRD corepressor complex to the methylated DNA (Hendrich *et al.*, 2001). A recent report, however, suggests that MBD2 and MBD3 do not coexist in the same complex but in two distinct complexes namely, MBD2/NuRD and MBD3/NuRD complexes (Le Guezennec *et al.*, 2006). Since MBD3 does not specifically interact with methylated DNA, it is unclear how the MBD3/NuRD complex is targeted to methylated DNA.

MBD2 null mice are viable and fertile; with normal imprinting patterns and methylation levels, however, these MBD2 knockout mice show impaired nurturing behaviour (Hendrich *et al.*, 2001). MBD2 deficiency is strongly correlated to suppression of intestinal tumorigenesis when crossed onto the Apc^{Min} background. $Mbd2^{-/-}Apc^{Min}$ mice live longer than Apc^{Min} controls showing a remarkable reduction in tumorigenesis (Sansom *et al.*, 2003). Even though the precise tumour suppression mechanism is unclear, it is interesting to note that many important negative regulators, for instance, secreted frizzled-related proteins (SFRPs) of the WNT pathway are transcriptionally repressed in colorectal cancer, implicating these negative regulators as potential targets of MBD2-mediated transcriptional repression (Suzuki *et al.*, 2004). This suggests a strong correlation between MBD2 and tumour suppressor genes. Subsequently identified target genes of MBD2 include the p161NK4A and p14ARF tumour-suppressor genes (Magdinier and Wolffe, 2001). This suggests the possibility of anti-cancer drug targeting of MBD2 in colorectal cancer.

MBD3, the smallest protein of the MBD family, consists of 285 amino acids with an acidic tail at the extreme C-terminus composed of 12 consecutive Glu residues (Figure 1-7). It shares a high amino acid sequence similarity to MBD2 with 71.1% overall amino acid identity (Hendrich and Bird, 1998). It is therefore believed that MBD2/3 represents the original methyl-CpG binding protein (Hendrich and Tweedie, 2003). The putative ancestral MBD2/3 protein is encoded by a single gene rather than distinct *MBD2* and *MBD3* genes as in vertebrates (Hendrich and Tweedie, 2003). The binding properties of MBD3 vary between species. Mammalian MBD3 does not specifically bind to methylated DNA due to the substitution of K43H and Y47F (Fraga *et al.*, 2003; Hendrich and Bird, 1998) whereas amphibian MBD3 binds strongly to methylated DNA (Hendrich and Bird, 1998). MBD3 is a component of the Mi2/NuRD chromatin-remodelling complex (Zhang *et al.*, 1999), however, the role of MBD3 in the complex remained unclear. Despite MBD2 and MBD3 amino acid sequences being closely related, gene targeting shows that the two proteins are not functionally redundant in mice, as MBD3 null mice die during early embryogenesis whereas MBD2 null are viable and fertile (Hendrich *et al.*, 2001).

1.4.5 MBD4

MBD4 shows the highest similarity to MeCP2 within the MBD domain. It consists of 580 amino acids and has two known functional domains: the MBD and thymine DNA glycosylase (TDG) domains are located at the C-terminal region (Hendrich and Bird, 1998). The MBD domain of MBD4 recognises symmetrical methylated or hemimethylated DNA with a strong preference for G-T mismatch (5'TpG3' complementary paired with 5'CpG3') and the TDG domain can efficiently remove thymine or uracil from the mismatches CpG site *in vitro* (Hendrich *et al.*, 1999). During DNA replication, 5'-methyl-cytosine has a higher tendency to mutate to thymine under spontaneous hydrolytic deamination (Figure 1-8), resulting in a G-T mismatch which can be repaired by thymine DNA glycosylase (Neddermann *et al.*, 1996; Wiebauer and Jiricny, 1989). In contrast, deamination of normal cytosine produces uracil, that can be efficiently detected and repaired by uracil DNA glycosylase (Nilsen *et al.*, 2001). The Human Gene Mutation Database (HGMD) reported that the amino acid mutation C-G to T-G mismatch accounts for over 20% of total single amino acid substitutions (Krawczak *et al.*, 1998).

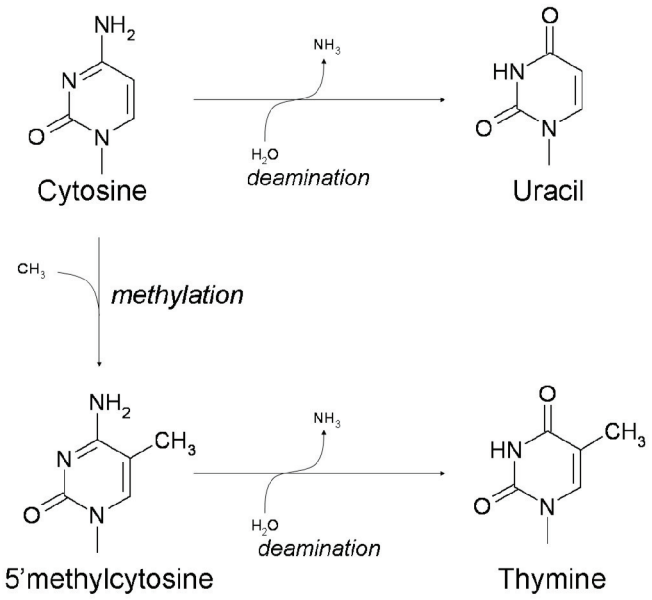


Figure 1-8 Hydrolytic deamination
Hydrolytic deamination leads to the formation of uracil whereas 5'methyl-cytosine produces thymine.

Unlike other MBD proteins, MBD4 is not associated with histone deacetylase activities although its transcriptional repression activity has been reported recently (Kondo *et al.*, 2005). Mutation of MBD4 has been found in TDG-deficient colorectal cancers, therefore, it is speculated that MBD4 can function as a *bona fide* tumour suppressor (Sansom *et al.*, 2007). MBD4-deficient mice are viable but exhibit a higher frequency of 5'methyl-cytosine mutation, in which, transition of methyl-CpG to TpG is 3-fold higher than wild-type mice (Millar *et al.*, 2002; Wong *et al.*, 2002). Deficiency of MBD4 in *mbd4*$^{-/-}$ mice also accelerates intestinal tumorigenesis in *Apc*$^{Min/+}$ background although spontaneous tumorigenesis was not observed (Millar *et al.*, 2002). These data strongly indicate that MBD4 functions as tumour suppressor via its DNA repair activity.

1.4.6 KAISO

Kaiso was first identified as a binding partner of the catenin protein p120 (Daniel and Reynolds, 1999). Kaiso is a bi-modal DNA binding protein that recognises methyl-CpG containing sequences and also a consensus sequence in the absence of methyl-CpG (Daniel *et al.*, 2002; Prokhortchouk *et al.*, 2001). Kaiso does not contain a classical MBD binding domain but interacts with methyl-CpG dinucleotides in a

context of mCpGmCpG via a zinc finger motif located at the C-terminus (Figure 1-7). It is capable of repressing transcription in a methylation dependent manner (Prokhortchouk *et al.*, 2001). In a nucleotide sequence recognition mode, the Kaiso zinc finger motif associates with the CTGCNA sequence motif (Daniel *et al.*, 2002). Depletion of Kaiso in *Xenopus* embryos leads to premature gene activation, abnormal gastrulation and early embryonic lethality (Kim *et al.*, 2002; Ruzov *et al.*, 2004). Repression by Kaiso is mediated through its stable association with the histone deacetylase-containing corepressor complex (NCoR) to a methylated site in the genome (Yoon *et al.*, 2003). Kaiso null mice are viable and fertile with no abnormalities of gene development and expression. However, a hybrid of kaiso null mice and tumour-susceptible $Apc^{Min/+}$ mice delayed the onset of intestinal tumourigenesis (Prokhortchouk *et al.*, 2006). Kaiso was also found to be upregulated in murine intestinal tumours and was expressed in human colon cancers (Prokhortchouk *et al.*, 2006). Therefore, kaiso represents a potential target for cancer therapy.

1.4.7 Structure of MBD domains

Solution structures of unliganded MBDs of MBD1 (Ohki *et al.*, 1999) and MeCP2 (Wakefield *et al.*, 1999) and of a DNA-bound MBD of MBD1 (Ohki *et al.*, 2001) have been reported. The structure of the MBD domain of MeCP2 is similar to the free MBD domain of MBD1 and in complex with DNA. As shown in Figure 1-9, the MBD structure is composed of a 4-stranded anti-parallel β-sheet and an α-helix. An extended long loop containing several positively charged residues connects β2 and β3. Several residues located at the C-terminal end of β2, loop L1 and the N-terminal end of β3 of MeCP2 MBD undergo chemical shift changes upon addition of methyl-CpG containing DNA, suggesting that this flexible loop is involved in DNA binding (Wakefield *et al.*, 1999). This is in agreement with the observation of the NMR structure of MBD1 MBD domain in complex with a 12 bp palindromic DNA containing a central methyl-CpG (Ohki *et al.*, 2001). Ohki and coworkers also argued that the methyl groups of methyl-CpG are recognised by a hydrophobic patch comprising of V20, R22, Y34, R44 and S45 (corresponding to K109, R111, Y123, R133 and S134 of MeCP2) (Ohki *et al.*, 2001). In addition, Wakefield et al (1999) reported that some residues located at loop L1 (G114 and A117), β3 (A131), loop L2

(R133), α1 (K135, V136), and the β-turn at the C-terminal region (F157, T160 and R162) showed amide proton chemical shift changes upon addition of methylated DNA.

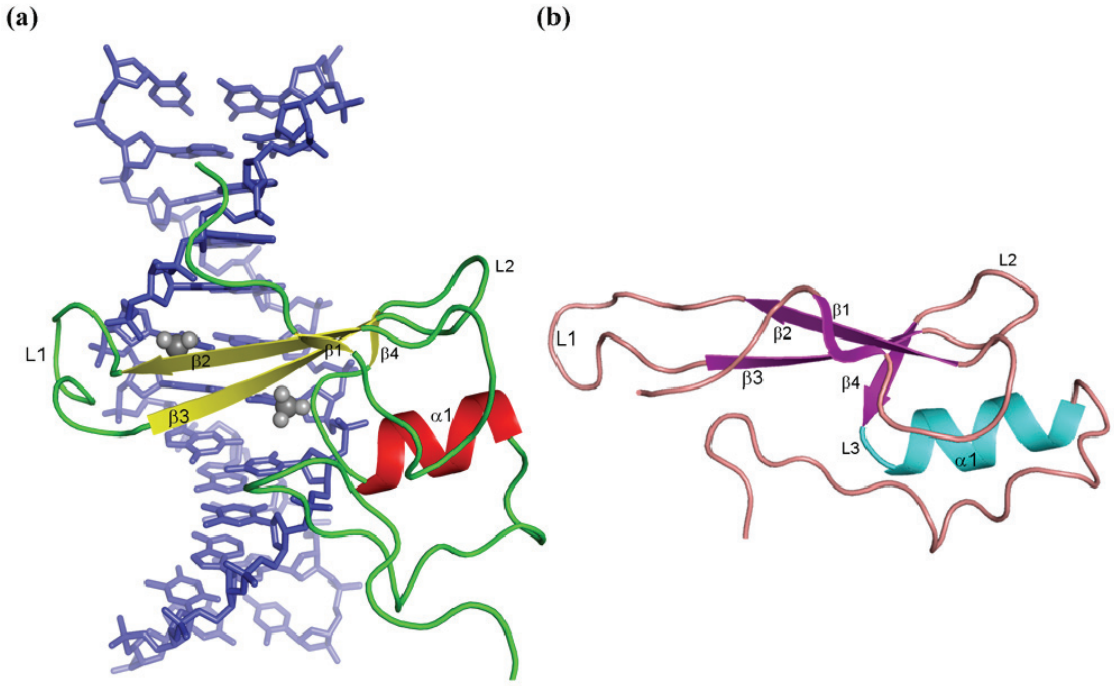


Figure 1-9: Solution structure of DNA bound MBD1 and unliganded MeCP2
(a) MBD domain of MBD1 (green, yellow and red) complexed with methylated 12 bp polindromic DNA (blue) (pdb id 1IG4; Ohki et al 2001). Methyl groups of methyl-CpG are represented by grey spheres. (b) Unliganded MeCP2 MBD domain (pdb id 1QK9; Wakefield et al 1999).

In the NMR structure of MBD1 MBD-DNA complex the following conclusions were drawn: The two methyl groups are separately packed into two hydrophobic pockets formed by the hydrophobic side-chains of Val20, Tyr34 and the aliphatic side chain of Arg22; and an aliphatic portion of Arg44 and Ser45 residues, respectively (Ohki *et al.*, 2001). However, subsequent work described in this thesis shows that the methyl-CpG is recognised through specific hydration at the major groove of the methylated-DNA. The details of the new findings are presented in Chapter 5 in this thesis and a significant portion has been published in Ho *et al.* (2008).

1.5 MECP2

1.5.1 Architecture of MeCP2

MeCP2 was originally identified in rat brain nuclear extract due to its ability to recognise single methyl-CpG dinucleotide (Lewis *et al.*, 1992) and an AT run adjacent to the methyl-CpG is an additional requirement to enhance maximum binding (Klose *et al.*, 2005). The full length protein contains 486 amino acids and is divided into several domains (Figure 1-10). The MBD domain of MeCP2 containing 85 amino acids located at amino acids 78-163. *In vitro* footprinting analysis indicates that 12 nucleotides surrounding a methyl-CpG pair is protected from DNase I activity, with an approximate dissociation constant of 10^{-9}M (Nan *et al.*, 1993). MeCP2 also contains a central TRD (amino acids 207-310) domain (Nan *et al.*, 1997) that interacts with various co-repressors complexes (Jones *et al.*, 1998; Kokura *et al.*, 2001; Nan *et al.*, 1998) and a nuclear localisation signal (NLS) (Nan *et al.*, 1996). In addition, two AT hooks, which are believed to interact with AT rich DNA, were identified at amino acid positions 185-192 and 265-277. These AT hook motifs are similar to that of the HMGA1 protein (Banks *et al.*, 1999; Dragan *et al.*, 2003). However, the AT hook binding specificity of MeCP2 has not been well characterised although Klose et al. reported that the AT hook domain had no effect on DNA binding (Klose *et al.*, 2005). MeCP2 forms a complex with methyl-CpG nucleosomal DNA exposed at the major groove via the MBD domain (Chandler *et al.*, 1999), indicating that accessibility of MeCP2 to nucleosomes is uninhibited. The C-terminal region is also required to facilitate the DNA-protein binding and contains a WW domain which is predicted to be involved in protein-protein interaction (Buschdorf and Stratling, 2004). The functional roles of the N-terminal (amino acids 1-77) and the C-terminal region containing 7x His tag and a Pro rich domain remain unknown. The expression of MeCP2 is from one locus located at the X-chromosome but alternative splicing gives rise to two isoforms (MeCP2α and MeCP2β) which differ only at their N-terminus (Kriaucionis and Bird, 2004; Mnatzakanian *et al.*, 2004). The expression efficiency of MeCP2 is different in various tissues although both isoforms are highly expressed in the brain; the roles of these alternative splice variants in normal MeCP2 function remain unknown.

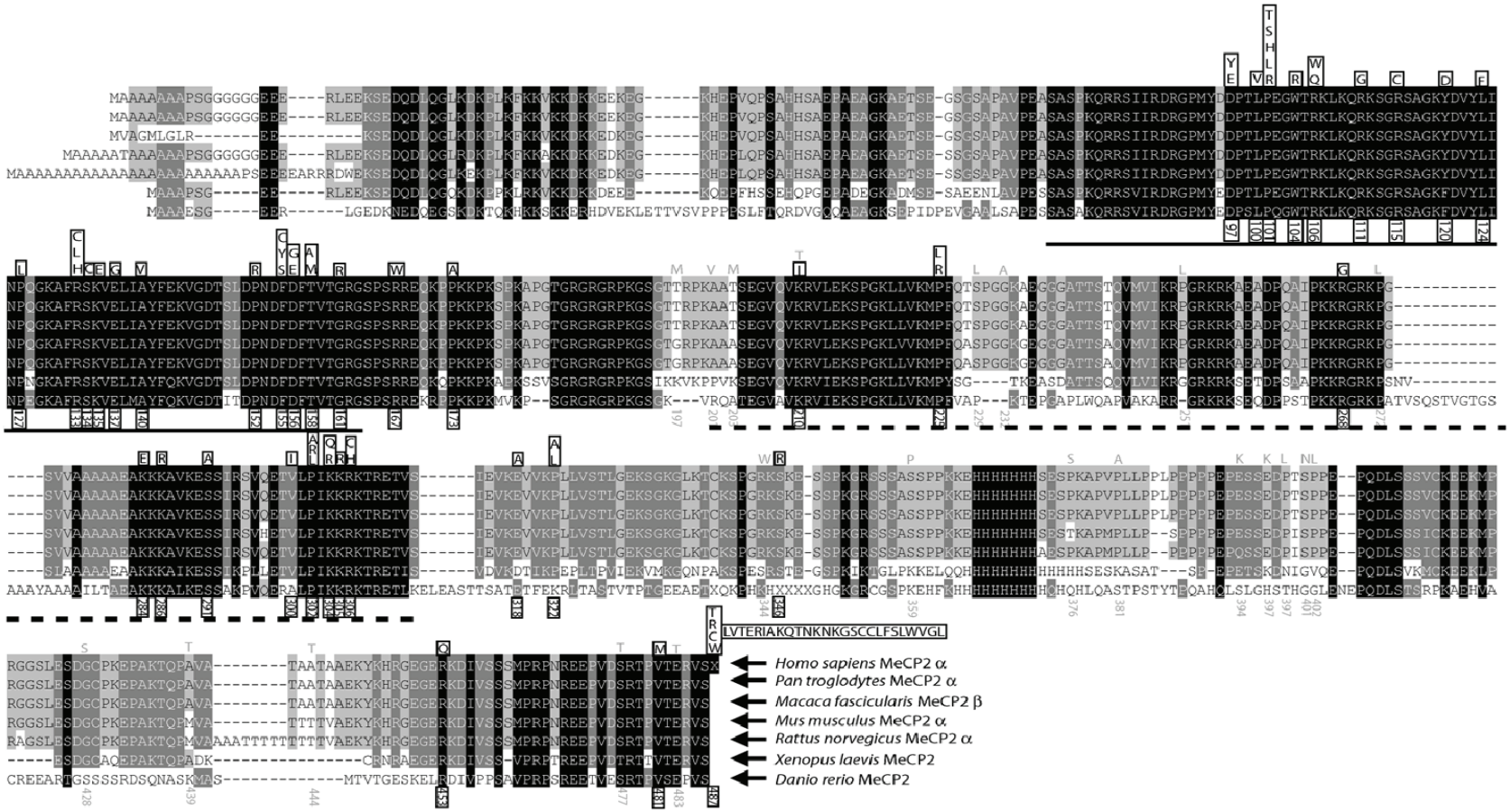


Figure 1-10 Alignment of MeCP2 proteins

The MBD and TRD domains are represented by the solid and dotted lines, respectively, under the amino acid sequence. The disease and non-disease causing mutations are indicated as black (boxed) and gray text, respectively, above the sequence alignment. The amino acid sequence below the sequence alignment corresponding to human MeCP2 sequence (alignment was provided by Dr. Skirmantas Kriaucionis).

1.5.2 MeCP2 Binding Partners

MeCP2 is known to recognise single methyl-CpG dinucleotides in DNA via its MBD domain; its TRD domain is used to recruit transcriptional repressor complexes and chromatin remodelling proteins (Jones *et al.*, 1998; Nan *et al.*, 1997). The TRD domain of MeCP2 is able to perform long-range transcriptional repression of up to 2kb from the mRNA initiation site (Nan *et al.*, 1997), and associates with the transcriptional repressor mSin3A and histone deacetylases (Jones *et al.*, 1998; Nan *et al.*, 1998). Gene silencing by MeCP2 is relieved by inhibition of histone deacetylase activity, which causes transcriptional derepression and chromatin re-modification (Jones *et al.*, 1998; Nan *et al.*, 1998). However, nuclear extracted native MeCP2 complexes demonstrated that the Sin3A co-repressor complex does not associate stably with MeCP2 (Klose and Bird, 2004), suggesting that MeCP2 is a weak or transient component of the complex and the protein might associate with a diverse range of cofactors. Other complexes have been gradually identified as MeCP2 binding partners in regulating transcriptional repression. Among them, the TRD of MeCP2 was reported to interact with transcription factor TFIIB by interfering with the assembly of the basal transcription machinery (Kaludov and Wolffe, 2000). MeCP2 was shown to recruit histone methyltransferase SUV39H1 which leads to the H3-K9 methylation in the differential methylated domain of *H19* gene, linking the repressive functions of DNA methylation and histone methylation (Fuks *et al.*, 2003a). MeCP2 also interacts with DNMT1 in a Sin3A/HDAC independent manner, which indicates that MeCP2 might be involved in guiding DNMT1, which lacks an MBD domain, to the hemimethylated CpG site at the DNA replication fork (Kimura and Shiota, 2003). Recently, MeCP2 was reported to interact with RNA binding protein YB1 (Y-box binding protein 1) which suggests a new role for of MeCP2 in regulating RNA alternative splicing (Young *et al.*, 2005). In addition to the above, MeCP2 interacts with various other factors in regulating gene transcription including c-ski, Brahma, SWI-SNF chromatin remodelling complex, group II WW domain of splicing FBPII and HYPC, PU1, Co-Rest, LANA, RNA and *Igfbp*3 (Buschdorf and Stratling, 2004; Harikrishnan *et al.*, 2005; Itoh *et al.*, 2007; Jeffery and Nakielny, 2004; Kokura *et al.*, 2001; Krithivas *et al.*, 2002; Lunyak *et al.*, 2002; Suzuki *et al.*, 2003). The mechanism and contribution of these factors to the function of MeCP2 remains to be investigated.

1.5.3 MeCP2 and Rett Syndrome

Rett syndrome (RTT) is a progressive neurodevelopmental disorder in early childhood and causes mental retardation in females, with a prevalence of 1 in 10,000 – 15,000 female births (Hagberg, 1985; Rett, 1966). Females with classical RTT develop phenotypically normal until 6-18 months, then gradually lose speech and purposeful hand use, and acquire microcephaly, dementia, seizures, autism, ataxia, intermittent hyperventilation and stereotypic hand movements (Hagberg *et al.*, 1983). After the initial regression, RTT patients usually stabilise and survive into adulthood. As RTT occurs exclusively in females, it has been suggested that the disease is an X-linked mutation with lethality in males due to severe encephalopathy (Bienvenu and Chelly, 2006).

Genetic mapping revealed that the disease region responsible for rare familial RTT located on Xq28 (Amir *et al.*, 2000) and candidate screenings identified mutations in the *MECP2* gene as the cause for most RTT cases (Amir *et al.*, 1999). *MECP2* mutations account for more than 95% of sporadic cases of classical RTT in females (Bienvenu and Chelly, 2006). Databases of disease-causing *MECP2* mutations and some benign polymorphisms have been compiled (see databases at the University of Edinburgh; www.mecp2.org.uk and RettBASE of the Children's Hospital at Westmead, Australia; www.mecp2.chw.edu.au). The compilation reveals that disease and non-disease causing missense mutations occur throughout the *MECP2* gene with more than half of all RTT causing mutations clustered within the MBD domain of MeCP2. Deletion/ insertion mutations also span the polypeptide but predominantly at the C-terminal region. Other forms of mutation are silent mutations, nonsense mutations, and miscellaneous mutations. About 64% of all mutations are clustered in 8 hotspots of RTT mutations; R106, R133, T158, R168, R255, R270, R294 and R306 (Figure 1-10); seven of them bearing CpG in their codon (the codon for Arg) (Kriaucionis and Bird, 2003). Therefore, it is likely that the unrepaired deamination of methyl-CpG causes those mutations (Cooper and Youssoufian, 1988). MeCP2 mutants display an abnormal DNA binding profile which eventually leads to over or underexpression of the target genes. A recent report suggests that RTT-like phenotypes are reversible by reinstalling MeCP2 expression in a mouse model (Guy *et al.*, 2007), indicating that the defective neurons are, in fact, viable.

1.5.4 MeCP2 target genes

MeCP2 is involved in methylation specific transcriptional repression and this observation has led to the hypothesis that MeCP2 deficiency in RTT causes global gene dysregulation. To test this hypothesis, transcriptional profiling using microarray analyses with *MECP2*-deficient mouse brain (Tudor *et al.*, 2002), RTT patient cell lines (Traynor *et al.*, 2002), post-mortem RTT brain (Colantuoni *et al.*, 2001) and recently cerebellum MeCP2 null mice (Venter *et al.*, 2001) has revealed no dramatic changes in genome wide transcription, indicating that MeCP2 does not function as a global gene repressor. One of the potential caveats is because MeCP2 regulates a different set of genes in different cell types and therefore the heterogeneity of the brain tissue would significantly mask cell-type-specific gene expression changes that may result from deficiency of MeCP2.

Several recent successful identifications of MeCP2 target genes relied on a candidate gene approach. By comparing gene expression profiling of $MeCP2^{+/y}$ and $MeCP2^{-/y}$ mouse cortical culture which is more homogeneous than whole brain, the *Brain Derived Neurotrophic Factor* (*BDNF*) gene was identified as the first *bona fide* mammalian endogenous target gene of MeCP2 (Chen *et al.*, 2003; Martinowich *et al.*, 2003). BDNF is known to play important roles in normal brain development and in learning and memory and these capabilities are disrupted in RTT patients (Chadwick and Wade, 2007). MeCP2 selectively binds to the *BDNF* promoter III and represses BDNF expression (Chen *et al.*, 2003). The *BDNF* gene is activated upon calcium-dependent phosphorylation at Ser421 which subsequently releases the phosphorylated MeCP2 (Chen *et al.*, 2003; Zhou *et al.*, 2006). However, a contradicting result was reported that the BDNF expression level was reduced in MeCP2 null mice (Chang *et al.*, 2006). A transgenic approach which allowed conditional BDNF overexpression in MeCP2 knockout mice relieved RTT symptoms observed in MeCP2 mutants (Chang *et al.*, 2006). This finding provides the first functional interaction between BDNF and MeCP2 in RTT disease progression. However, it is still unclear how MeCP2 and BDNF cooperatively regulate neuron activities.

In an amphibian model, the neuronal repressor *Hairy2a* gene was identified as an MeCP2 target gene (Stancheva *et al.*, 2003). In this model, Stancheva and colleagues used antisense morpholino oligonucleotide injection to knock-out the *MECP2* expression. In the MeCP2 deficient frogs, significant developmental defects such as reduced dorsal axis and abnormal head structure have been observed (Stancheva *et al.*, 2003). Disruption of MeCP2 activity also affect the primary neuron patterning during neuron differentiation indicating that lack of MeCP2 causes neurogenesis problems. The dysregulation of gene expression was analysed in the Delta/Notch signalling pathway, which is known as the key pathway in early neuronal development. MeCP2 deficiency leads to increased expression of *Hairy2a* gene. Mechanistically, MeCP2 binds to methyl-CpG of the *Hairy2a* promoter, where, MeCP2 interaction with SMRT complex via Sin3A represses *Hairy2a* gene. On activation of Delta/Notch signalling pathway, the repression complexes containing MeCP2, SMRT, Sin3A and HDACs leave the methyl-CpG sites.

Several other methods have been used to identify primary targets of MeCP2. *DLX5* was identified by cloning fragments from MeCP2 ChIP (Horike *et al.*, 2005). Upregulation of serum glucocorticoid-inducible kinase 1 (*sgk*) and FK506 binding protein 5 (*Fkbp5*) in MeCP2 deficient mouse brain, suggesting MeCP2 as a modulator of glucocorticoid-inducible gene expression (Nuber *et al.*, 2005). Microarray analysis of transcriptional changes during maturation differentiation identified inhibitors of differentiation (ID1-4) genes as a neuronal target of MeCP2 as both *MECP2*-null mice and RTT brains show elevated ID proteins (Peddada *et al.*, 2006). FXYD1, which encodes a transmembrane modulator of Na^+, K^+-ATPase activity, is enhanced in frontal cortex (FC) neurons of RTT patients and MECP2-null mice (Deng *et al.*, 2007). The target genes identified in these studies show dysregulation due to MeCP2 deficiency but none of the target genes could completely explain the primary neurodevelopment defect observed in RTT patients.

1.5.5 Methylation dependent and independent Mechanisms

A series of studies have identified MeCP2 endogenous target genes that function in two different pathways that might contribute to RTT syndrome. The classical pathway involves recognition of methylated DNA at a target promoter region via the protein

MBD domain. The second pathway however might not involve methyl-CpG recognition but establishes a silent chromatin structure through formation of a chromatin loop. Figure 1-11 illustrates how MeCP2 functions in methylation and non-methylation dependent manners.

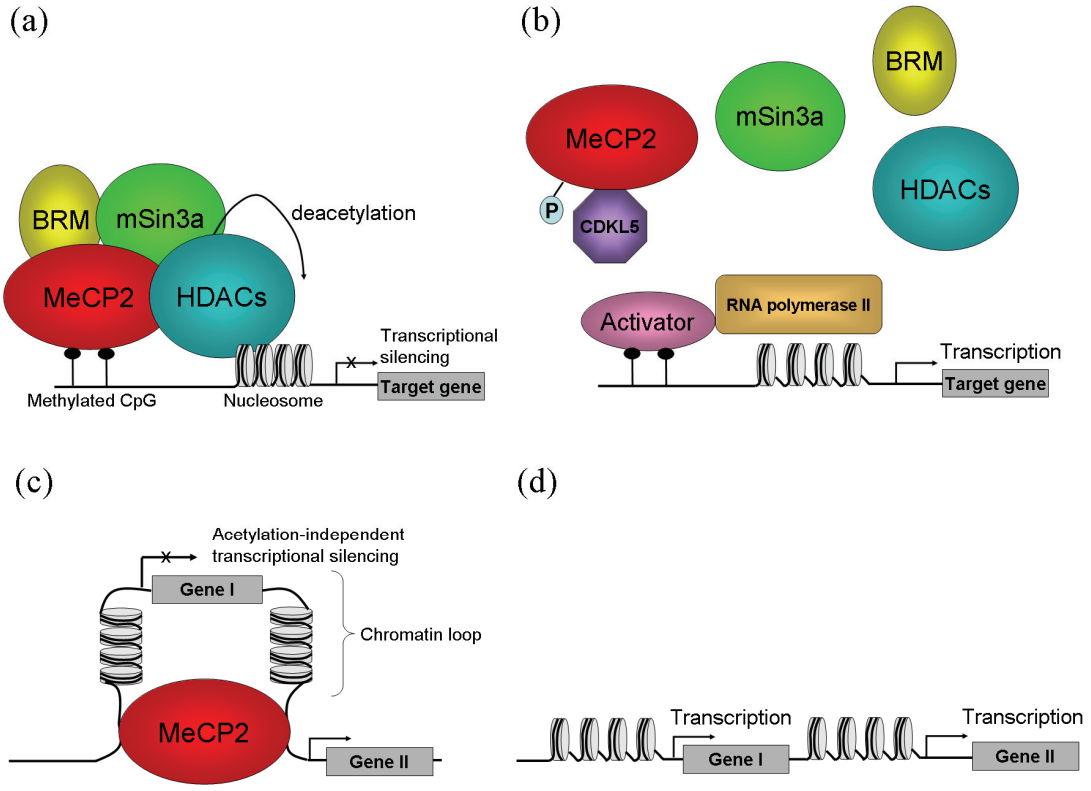


Figure 1-11: MeCP2 regulation of chromatin remodelling and transcriptional
(a) MeCP2 binds methylated DNA through the MBD domain and the TRD domain recruits chromatin-remodelling complexes that contain mSin3A co-repressor, BRM and HDACs. The recruited HDACs acetylates histones H3 and H4, which causes chromatin condensation and subsequently, limiting the accessibility of transcriptional machinery.
(b) Membrane depolarisation followed by calcium influx leads to activation of the target gene. Calcium dependent phosphorylation releases the MeCP2 from its target site and disassembling the co-repressor complexes. The displacement of MeCP2 from its target site has been suggested from reduced CpG methylation at relevant CpG sequence (Martinowich *et al.*, 2003). Alternatively, the dissociation could be due to CDKL5 activity which is thought to bind and phosphorylate MeCP2 (Mari *et al.*, 2005).
(c) MeCP2 binds to its target DNA sequence and result in formation of a chromatin structure, followed by transcriptional silencing independently of DNA methylation.
(d) The absence of MeCP2 at the DNA methylation independent promoter leads to a relieved form of chromatin structure and rendering accessibility to transcriptional machinery (model adapted from Bienveno and Chelly, 2006).

In the methylation dependent model (Figure 1-11a and b), the MBD domain initially binds onto methyl-CpG sites and its TRD domain interacts with the co-repressor Sin3A complex which contains HDAC1/2. Deacetylation of histone H4 by HDACs remodels the chromatin structure, which becomes inaccessible to the transcriptional machinery (Bienvenu and Chelly, 2006). Various other transcriptional repression factors have been reported to associate with the TRD domain of MeCP2 in regulating transcriptional repression (see MeCP2 binding partners). However, the link between histone deacetylation and MeCP2-mediated transcription repression remains to be fully addressed. In addition, histone methylation is another important epigenetic development in regulating chromatin structure and activity (Kouzarides, 2002). MeCP2 has been reported to be involved in methylation of H3K9 on the *H19* gene, which is associated to gene silencing (Fuks *et al.*, 2003a; Fuks *et al.*, 2003b). Other complexes are believed to mediate the association of MeCP2 and H3K9 methylation. MeCP2 is also reported to interact with the repressor complex RCOR1. This complex binds to the SCN2A gene and represses transcription through H3K9 methylation, which is catalysed by the histone lysine methyltransferase SUV39H1. In most cases, however, the role of MeCP2 remains elusive.

In 2003, a novel mechanism of MeCP2 in transcriptional repression independent of DNA methylation was proposed (Figure 1-11c and d). *In vitro* biochemical studies and electron microscopy demonstrated that the complex of MeCP2 and unmethylated nucleosomal arrays resulted in extensively condensed ellipsoidal particles and oligomeric suprastructures (Georgel *et al.*, 2003). Surprisingly, analysis with RTT mutants (R168X and R133C) indicates that chromatin condensation is dispensable of the MBD domain (Georgel *et al.*, 2003). These data suggest that MeCP2 binding has an enormous impact on chromatin reorganisation in a methylation independent manner. This result is in agreement with the finding that *Dlx5-Dlx6* are targeted imprinted genes of MeCP2. MeCP2 mediated the silent chromatin-derived 11 kbp chromatin loop at *Dlx5-Dlx6* whereas this loop is absent from MeCP2 null brain (Horike *et al.*, 2005). Interestingly, this finding did not identify any sequences in the region of identified MeCP2 binding sites that were differentially methylated on the two parental alleles (Horike *et al.*, 2005).

1.6 PROJECT AIMS

This study concerns the molecular basis of the DNA binding specificity of the transcriptional repressor MeCP2, which is essential for maintenance of neuronal functions (Guy *et al.*, 2007; Kishi and Macklis, 2004). The aims of the study include:

1. Characterisation of MeCP2 binding with various DNA fragments selected from SELEX experiments (Klose *et al.*, 2005) using gel retardation assays. In particular the role of AT runs adjacent to the methyl-CpG will be investigated.
2. Co-crystallisation of MeCP2 MBD in complex with a methylated DNA fragment.
3. Structural determination of the MeCP2 MBD-DNA complex using X-ray crystallography.
4. Identification of the structural role played by several key residues in RTT mutation.

Chapter 3 presents preliminary characterisations of various MeCP2 constructs complexed with methylated DNA and the selection of complexes for initial co-crystallisation trials. Chapter 4 describes the initial crystallisation screening and optimisation of MeCP2 MBD in complex with various lengths of DNA containing a methyl-CpG and an AT run at the centre of the duplexes. This process established the crystallisation condition for the DNA-protein complex of an MeCP2 MBD domain and a 20 bp overhanging methylated DNA. This chapter also explains the X-ray crystallographic strategy used in solving the X-ray structure of MeCP2 MBD domain in complex with a methylated DNA. Chapter 5 describes the details of structural analysis using various programmes. Chapter 6 presents various gel shift assays which highlight the importance of several residues involving in methylated DNA binding. Significant parts from Chapter 4, 5 and 6 have been published in Ho *et al*. (2008).

CHAPTER 2. MACROMOLECULAR CRYSTALLOGRAPHY

2.1 INTRODUCTION

Macromolecular crystallography is a technique used to study the structure of biological molecules such as proteins, nucleic acids and viruses to atomic resolution. High resolution X-ray crystal structures help to elucidate the details of mechanism of macromolecules in living cells and organisms. Most macromolecules can be crystallised under regulated conditions. In this state, macromolecules are arranged in a regular three-dimensional arrangement. Atoms in equivalent positions can diffract X-rays to produce diffraction spots. The X-ray structure of the macromolecule can be determined by analysing the spot intensities and position. This chapter provides an introduction to the basic concept of macromolecular crystallisation and crystallography used in this study.

2.2 MACROMOLECULAR CRYSTALLISATION

Solubility of macromolecules such as proteins and DNA in aqueous solutions is a function of parameters such as pH, temperature, ionic strength and macromolecular concentration. Varying these parameters can bring the macromolecule to a thermodynamically metastable state known as supersaturation, in which, the amount of macromolecule present in the solution is higher than that of the macromolecular solubility. In the supersaturation state, the macromolecules can evolve kinetically into two different states; aggregates or nuclei; depending on the molecular interaction control (Durbin and Feher, 1996). To promote crystal growth for structural determination, the supersaturated state is controlled in a way to promote limited nucleation that allows crystal growth, but not over nucleation, which results in many microcrystals or amorphous precipitate. A crystallisation phase diagram (Figure 2-1) illustrates four distinct areas present in the supersaturation state: (i) at the highest supersaturation region, protein precipitation will occur; (ii) at the moderate supersaturation area, spontaneous nucleation will take place; (iii) at the area of lower supersaturation just below the nucleation zone, crystals are stable and may grow but no further nucleation takes place; and (iv) at the under saturation region, the macromolecule is fully dissolved and crystallisation will never occur (Chayen, 2004;

Chayen, 2005). One of the key steps in growing X-ray diffractable crystals is control of nucleation. The ideal strategy for growing large crystals is to allow restricted nucleation at the boundary between the nucleation and metastable zones, which results in a few nuclei and subsequently reduces macromolecular concentration. A reduction in macromolecular concentration transforms the nucleation zone to the metastable zone across the supersolubility curve. The metastable condition allows crystals to grow without additional nucleation.

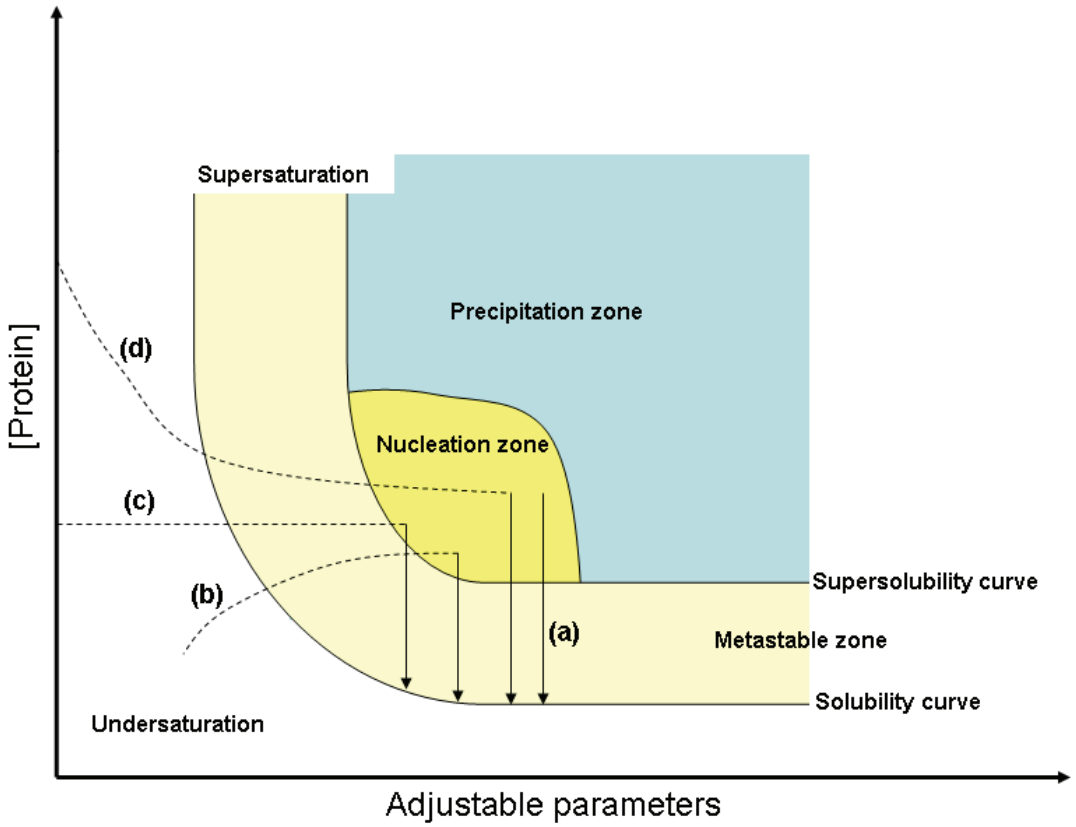


Figure 2-1 Schematic illustration of a protein crystallisation phase diagram
The adjustable parameters can be precipitant and additive concentrations, pH or temperature. Route to nucleation and metastable zones of four major crystallisation methods are labelled as (a) batch method; (b) vapour diffusion; (c) dialysis and (d) free interface diffusion (diagram adapted from Chayen 2004).

In practice, finding an appropriate condition for macromolecules to crystallise could be time and material consuming. Crystallisation experiments are based mainly on trial and error at the early stages of screening. However, there are two screening approaches for finding initial crystallisation conditions (McPherson, 2004). The first approach is a systematic variation of variables such as precipitant type and concentration; macromolecule concentration, pH, temperature and other parameters;

to search for initial crystallisation condition. The second is what termed as a 'shotgun approach' which is a sparse matrix screen utilising a wide range of selected conditions that have been proved successful in crystallising other macromolecules. The screen contains a wide combination of precipitants, salt, pH, ions and physical parameters most frequently reported in the literature. The Natrix Screen (Hampton Research) containing 48 conditions is an example of sparse matrix formulations selected from known crystallisation condition for nucleic acid-protein complexes. Once the initial hit is found, the crystallisation condition is optimised carefully in a relatively narrow range of parameters to yield an X-ray diffractable single crystal. Recent advancement in robotic automation and high-throughput screening allows a huge number of conditions to be tested within a short period of time, which requires only a small amount of sample.

Preliminary studies prior to initial crystallisation screening are required to gain knowledge of biochemical properties of the macromolecules. Information such as solubility, macromolecular size, multimerisation state, requirement of cofactor and temperature stability can be acquired during macromolecule purification. Light scattering and circular dichroism analysis can estimate monodispersity and secondary structure of the protein. Lastly, analysing related macromolecular structures can guide a rational 'tailoring' of the protein of interest by modifying the protein surface by mutagenesis or truncating the flexible loops (Derewenda, 2004). In DNA-protein co-crystallisation, the ratio of DNA to protein in complex form must be determined experimentally using various methods such as band shift assay or gel filtration analysis. This knowledge could significantly narrow down the search for early crystallisation conditions.

2.2.1 Crystallisation methods

The strategy used in crystallisation is to bring the macromolecule/solvent system to achieve supersaturation by reducing the solubility of a macromolecule by means of modifying the properties of the solvent or the character of the biomolecules. Several systems have been invented to achieve the supersaturated state of the macromolecular solution. These include bulk crystallisation, batch methods, dialysis, seeding, free interface diffusion, vapour diffusion and temperature induced methods (McPherson,

1999). Currently, the most popular method in crystallisation is hanging drop vapour diffusion.

2.2.1.1 Vapour diffusion

There are two common procedures in setting up vapour diffusion crystallisation. The mother liquor which contains the macromolecules and precipitant solution is either suspended (hanging drop) or supported (sitting drop) by some surfaces (Figure 2-2). In these approaches, the protein solution drop on the surface, usually 0.5 to 20 μl, is spatially separated from the reservoir, which is several magnitudes larger in volume. The precipitant concentration in the drop is generally lower (half, in general practice) than that of the reservoir solution. To achieve equilibrium in a sealed system, at a chosen temperature, the water or volatile material is evaporated from the drop until equilibration. The reservoir concentration is essentially unchanged. The drop volume decreases slowly which increases the macromolecular concentration. Generally, the precipitant concentration in the drop is half diluted by mixing with an equal volume of sample. The initial concentrations of the precipitant, macromolecule and additive are refined until supersaturation is reached upon equilibration.

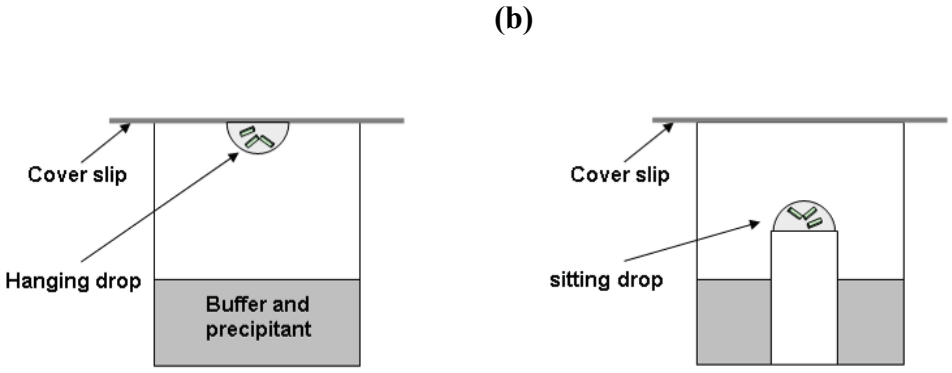


Figure 2-2 Vapour diffusion methods
The mother liquor is (a) suspended on a cover slip in a hanging drop preparation and (b) supported on a sitting drop preparation.

2.2.1.2 Microbatch

Batch crystallisation is the simplest method in crystallising a protein or nucleic acid. The crystallisation mixture is prepared by direct mixing of an undersaturated macromolecule solution with a precipitant solution. The precipitant alters the protein solubility to create an immediate supersaturation state. If the chemical and physical

parameters have been carefully chosen, the solid state will be crystalline rather then amorphous. In the microbatch method, a 96-well microtiterplate is used as crystallisation platform by mixing the macromolecule and precipitant solutions under a layer of oil. The thickness of the oil can be used to control evaporation of water thus regulating nucleation which can be ceased later by increasing the oil depth (Chayen and Saridakis, 2002). The microbatch method has been adapted for automated high-throughput screening experiments which consuming less than 500nl sample per well (Luft *et al.*, 2001).

2.2.1.3 Crystallisation by Dialysis

Dialysis is a process of separating molecules in solution according to size through the use of semipermeable membrane containing pores of less than macromolecular dimension. In crystallisation by dialysis, the biological macromolecule is separated from a large volume of precipitant solution by a semipermeable membrane which allow small molecules such as ions, additives and buffer to pass through but not the larger size biological macromolecules. The kinetics of equilibrium depends on the membrane cut-off, the ratio of the precipitant concentration inside and outside the biological macromolecule chamber, temperature and the geometry of the cell. The simplest technique is to use a dialysis bag but a large volume of sample is required. Alternatively, microdialysis cells such as Zeppenzauer cells and dialysis buttons (commercially available from Hampton Research) can be used. These microdialysis cells have been adapted to accommodate a micro-volume of sample.

2.2.1.4 Cryocrystallography

Diffraction data are normally collected from crystal continuously cooled in a nitrogen stream. In theory, lowering the temperature should increase molecular order in the crystal and improve diffraction. However, freezing macromolecular crystals in liquid nitrogen occasionally results in ice formation which potentially damages the macromolecular crystal or interferes with data processing even though current software can handle ice-ring interference effectively. To overcome this problem, cryocrystallography was developed and cryoprotectants such as glycerol and low molecular weight PEG are introduced into the crystallisation solution or soaked into the crystal to prevent ice formation. Other advantages of cryocrystallography include

cryoprotection of macromolecular crystals against radiation damage particularly with intense synchrotron radiation sources. A comprehensive review and current advancements of this technique can be found in Garman and Owen (2006).

2.3 CRYSTALS AND SYMMETRY

A crystal is a solid object composed of an orderly three-dimensional array of molecules, held together by non-covalent interactions, which can be described as a three dimensional arrangement of lattice points is known as a space lattice. The lattice points selected as the vertices of the unit cell are chosen in a way that is consistent with the highest possible symmetry with the shortest cell edges. The edges are defined by 3 vectors **a**, **b** and **c** and the angles between these defined as α, β and γ (Figure 2-3). In a unit cell, operation of the internal symmetry elements can further define the asymmetric unit. The asymmetric unit is the repeating object which is related to all other identical objects within the unit cell.

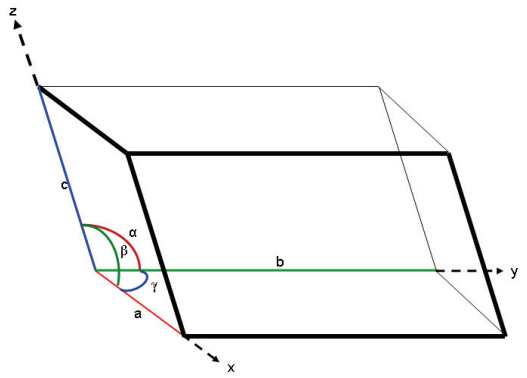


Figure 2-3 General unit cells
A unit cell is defined by three edges **a**, **b** and **c** and angles α, β and γ (diagram adapted from Blow 2002).

On the basis of symmetry, crystals can be grouped into 7 crystal systems (Table 2-1). The simplest crystal system is triclinic where $a \neq b \neq c$ and $\alpha \neq \beta \neq \gamma$. If $a \neq b \neq c$ and $\alpha = \gamma = 90^{\circ}$, and $\beta > 90^{\circ}$, then the cell is monoclinic. In order of increasing symmetry, they are triclinic, monoclinic, orthorhombic, trigonal, tetragonal, hexagonal, and cubic. The space group describes the lattice type and the internal symmetry elements of the unit cell. Primitive lattice (P) contains 1/8 lattice point at each vertex of the cell and therefore contains a total of one lattice point per unit cell. An internal (I) or body-centred lattice contains two lattice points per unit cell, one in the centre of the cell and

1/8 at each corner of the cell. A face-centred lattice (F) contains additional lattice points at the centre of all faces of the unit cell and thus comprises 4 lattice points. In a monoclinic crystal system, if a lattice point is placed on one of the faces; A, B or C; then new lattice types can be created. However, by changing cell orientation, A and B faced-centred lattices are equivalent to C-centred lattice. By combining the crystal systems and lattice types, only 14 Bravais (space) lattices are allowed (Table 2-1). To define the internal symmetry of the unit cell, three symmetry operations and elements are needed. Translation is a movement along the unit cell axes usually by fractions of the cell edges. Rotation is a movement of an object about an axis which can be expressed as x-fold rotation and the rotation angle is 360˚/x. The symmetry operation of a screw axis results from a combination of rotation and followed by translation parallel to the rotational axis. The combination of the various symmetry operators gives rise to 230 possible space groups. Because biological molecules are enantiomorphic, only 65 enantiomorphic space groups are considered for macromolecular crystals.

Table 2-1 The seven crystal systems

Crystal systems	Possible Bravais lattices	Minimum symmetry requirement	Constraints on axial length	Constraints on interaxial angles
Triclinic	P	None	$a \neq b \neq c$	$\alpha \neq \beta \neq \gamma$
Monoclinic	P, C	One 2-fold axis	$a \neq b \neq c$	$\alpha = \gamma = 90^\circ \neq \beta$
Orthorhombic	P, C, I, F	3 perpendicular 2-fold axis	$a \neq b \neq c$	$\alpha = \gamma = \beta = 90^\circ$
Trigonal	P (or R)	One 3-fold axis	$a = b \neq c$	$\alpha = \beta = 90^\circ, \gamma = 120$
Tetragonal	P, I	One 4-fold axis	$a = b \neq c$	$\alpha = \beta = \gamma = 90^\circ$
Hexagonal	P	One 6-fold axis	$a = b \neq c$	$\alpha = \beta = 90^\circ, \gamma = 120$
Cubic	P, I, F	Four 3-fold axis	$a = b = c$	$\alpha = \gamma = \beta = 90^\circ$

2.3.1 Miller indices

It is possible to draw families of planes that contain lattice points in the space lattices. These theoretically constructed planes intercept the unit cell axes fractionally, where the interception produces an integer number of equal segments for each axis a, b and c. The number of segments generated in each axis is called the Miller index of these planes and is represented by integer numbers *hkl*. For a given unit cell with axes a, b and c, if a family of planes parallel to **a** axis, cuts **b** axis by half and **c** axis into 3 equal segments, then the Miller index for these parallel planes is 023. Figure 2-4

illustrates that the Miller index defines the direction of these planes with respect to the unit cell orientation of the crystal. The interplanar distance is designated as d_{hkl}.

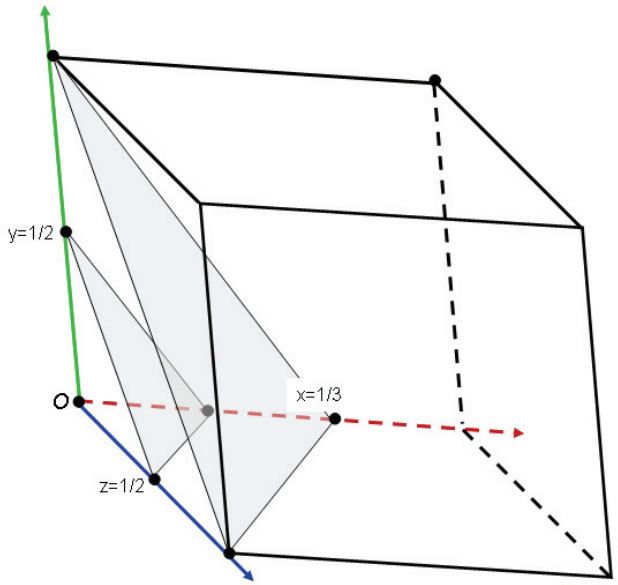


Figure 2-4 Planes in space lattice
Each family of planes is designated with a Miller index which can be used to define the direction of these planes correspond to the unit cell axes. The Miller index in this example is 322.

2.4 X-RAY AND DIFFRACTION

2.4.1 X-ray

X-rays are electromagnetic radiation of wavelengths 0.1-100Å (1 Å = 10^{-10} m) which can be produced when a metal is bombarded with electrons accelerated by an electric field. For macromolecular crystallography, the wavelength ranges between 0.5 to 1.6 Å are most suitable as these waves can penetrate biological samples and be scattered by the entire volume of a macromolecular crystal. X-rays are produced when a high energy electron collides with the target metal and displaces an electron from the lower orbital. This vacancy is filled by an electron dropped from a higher orbital and this transition process emits the excess energy in the form of X-ray photons. In crystallography laboratories, electrons are generated from a hot filament and are accelerated by a large (40kV) potential. The high speed electron then collides with the water-cooled anode (or rotating anode) of the target metal and promotes electronic transitions between orbitals resulting in X-ray radiation. Two of the most common anode elements are copper and molybdenum. When the electron from the inner shell

is displaced by accelerated electron, the electron from L shell of copper will drop and replace displaced K electron (K_α transition) and emit X-rays with wavelength of 1.54 Å. The initial beam however contains two wavelengths (K_α and K_β) which will produce overlapping reflections and causes difficulty in indexing. To overcome this problem, a monochromator is used to select the stronger wavelength and produces monochromatic X-rays.

In synchrotron radiation sources, electrons circulate at velocities near the speed of light in a giant ring (with circumference up to 1km). The electron is forced into a curved motion by powerful magnets in accelerators and additional bending is imposed by wigglers. The curving and bending of electrons causes emission of high energy X-ray (synchrotron radiation). The intensity of a synchrotron X-ray beam is several times stronger than the in-house X-ray sources using either X-ray sealed-tube or rotating anode sources. This feature allows data collection to be completed in short time and also permits data collection from small and weakly diffracting crystals. In addition, synchrotron radiation produces radiation with a continuum of wavelengths which allows a desired wavelength to be selected, which is particularly helpful in solving the phase problem.

2.4.2 X-ray diffraction

Diffraction occurs when an incident beam is scattered by an object into directions other than the original direction without changing its wavelength. In macromolecular crystallography, X-rays interact with electrons and set the electrons oscillating with the X-ray frequency and emit an X-ray photon with the same wavelength in a random direction. Unlike visible light, X-ray radiation cannot be focused by lenses. Reproducing the image requires the phases of the scattered wave as well as the experimental measured intensity. To 'visualise' an X-ray diffracted image, the phases must be estimated by calculation. The X-ray diffraction image is recorded on devices which can include: X-ray sensitive film, image plate or CCD detector. These area detectors can record the diffraction pattern, the positions and the intensities of each diffracted spot. During data collection, the crystal is mounted on the goniometer head which is rotated by the diffractometer through an angle which allows a complete or

high multiplicity dataset to be collected depending on structural determination strategies and space group.

2.4.3 Bragg's Law

As mentioned above, the X-rays diffracted from each atom are in random directions with an identical wavelength to the incident beam. Because the molecules are in a regular repeated lattice, the scattered rays will interfere constructively producing strong diffracted spots or destructively cancelling each other. This phenomenon can be explained by Bragg's Law using an equation:

$$2d_{hkl}\sin\theta = n\lambda \qquad \textbf{Equation 2-1}$$

where d_{hkl} is the distance between parallel planes (hkl), θ is the angle of diffraction, λ is the wavelength and n is an integer. Only in the condition when the difference in path length for rays diffracted from successive planes is equal to an integral number of wavelengths ($n\lambda$) will constructively produce a strong diffracted Bragg reflection. At other angles of incidence, diffracted waves from successive planes interfere destructively and no Bragg spot will be recorded. Each family of successive planes that satisfy the Bragg's Law produces a reflection. Each Bragg spot on the diffraction image result from a different set of Miller planes.

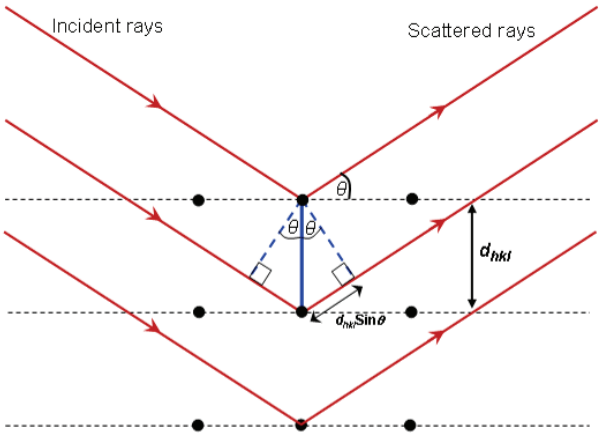


Figure 2-5 Bragg's Law

The reciprocal lattice dimension is inversely proportional to the real space dimension. The base vectors of the reciprocal space are named **a***, **b*** and **c*** which are perpendicular to the real space planes bc, ac and ab with magnitude proportional to 1/**a**, 1/**b** and 1/**c**, respectively. The relationship between the real space in a crystal and its space occupied by the reflections in reciprocal space can be described using the

Ewald sphere concept (Figure 2-6). The Ewald construction is a sphere, where the centre C is placed in a crystal, with radius equal to $1/\lambda$. The crystal lattice is represented by the reciprocal lattice, with its origin O on the Ewald sphere rotational axis where the direct beam ACO leaves. When a lattice point B touches the sphere, Bragg's Law is fulfilled and a reflection can be recorded. As a result, the diffraction of each family of planes (Miller indices *hkl*) that satisfy Bragg's Law will contribute to a reciprocal lattice point in reciprocal space. A vector (OB) is drawn perpendicular to each family of *hkl* planes with length of $1/d_{hkl}$, the points at the end of the vectors define the three-dimensional lattice known as the reciprocal lattice.

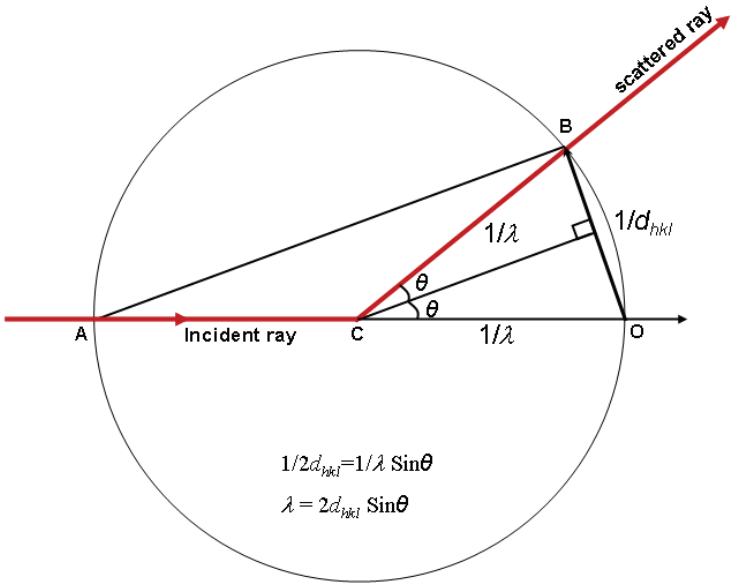


Figure 2-6 Ewald sphere construction
The reciprocal lattice point B is defined by the end point of vector OB with length of $1/d_{hkl}$. When the rotation of the crystal brings point B onto the surface of the sphere then ½ *OB* = *CB* $Sin\theta$ which is equivalent to $1/2d_{hkl} = 1/\lambda \sin\theta$ and re-arrangement of this equation proves the Bragg's law; $\lambda = 2d_{hkl} \sin\theta$.

2.5 DATA COLLECTION STRATEGY

The method of choice in data collection is the rotation method based on the principle that rotating a crystal around an axis should bring most of the reciprocal lattice points to the conditions that fulfil Bragg's law. A single image is recorded every time the crystal is rotated by a definite small angle for a given length of time. This is the oscillation angle of the diffraction images. The objective is to bring as many as possible reciprocal lattice points to the Bragg's condition in order to obtain a high completeness for the dataset. To collect good quality data with high completeness,

experimental parameters that should be considered in the data collection strategy are oscillation angle, total rotation range, crystal-to-detector distance, wavelength and exposure time. The strategy used also depends on crystal properties such as possible space group, crystal mosaicity, unit cell and maximum resolution. Therefore, two images at 90˚ apart are usually collected and analysed with data processing programmes in order to estimate the diffraction properties of the crystal. This preliminary information can be used to adjust the crystal-to-detector distance according to the extension of diffraction. Autoindexing of these images can predict the possible space group and the unit cell dimensions.

Two approaches can be used to select the rotational angle (angular width) of an individual reflection. The wide slicing (wide angle) method is based on collecting images with an angle greater than the mosaic spread of the crystal thus more fully-recorded reflections are recorded. Two-dimensional integration of reflections in each image can be performed with MOSFLM (Leslie, 1999) and HKL (Otwinowski and Minor, 1997). For the fine slicing (small angle) method, images are collected with the sliced angle narrower than the mosaic spread of the crystal or less than or equal to 0.5˚. A "thin" dataset has reflections spread over several images which enable 3-dimensional integration of reflections using programmes such as XDS and d*TREK (Pflugrath, 1999). The fine slicing method results in a better signal-to-noise ratio compared to the wide angle method. However, the fine slicing method is only advantageous if a detector has a very short readout time.

The setting of crystal-to-detector distance affects the signal-to-noise ratio in the recorded diffraction pattern. A long crystal-to-detector distance gives a better signal-to-noise ratio but may restrict resolution of the data and *vice versa*. Therefore, crystal-to-detector distance should be adjusted to match the maximum resolution limit which can be estimated by the 2 images separated by 90˚. In addition, the crystal-to-detector distance setting is also affected by the unit cell dimension. The highest resolution setting for a large unit cell dimension, such as a virus particle, would potentially lead to overlap reflection profiles. In some cases, the resolution is compromised for a full completeness dataset.

The beam wavelength is extremely important in collecting the anomalous signal for experimental phasing and therefore it must be appropriately optimised. This optimisation can be done at a station; such as stations BM14 at ESRF where the wavelength can be tuned according to the fluorescence spectrum of the anomalous scatterers in the crystal. Furthermore, the fluorescence spectrum also provides *f'* and *f''* values which are required for a MAD experiment. For native data collection, any value of the wavelength which ensures a high intensity of the beam can be used. Most of the synchrotron radiation sources utilise wavelengths below 1Å as this minimises the absorption of radiation by the crystal, the mother liquor and air scatter.

The total rotation angle is the most important factor that influences the completeness of data. In principle, collecting 180˚ of data will ensure a maximum completeness or 360˚ if an anomalous scatterer is present (Dauter, 1999). The analysis of the crystal symmetry in relation to the geometry of the rotation method allows a complete dataset to be collected in which unique reflections are measured at least once. Nevertheless, in cases where the crystal is vulnerable to radiation damage, particularly in 3rd and 4th generation synchrotron radiation sources, it would be beneficial to reach high completeness as soon as possible using an optimal data collection strategy. For successful usage of anomalous data in MAD and SAD phasing, the important aspect of the data is its quality in terms of completeness of the Friedel pair reflections and high multiplicity in order to increase the signal-to-noise ratio of a weak anomalous signal. However, any reciprocal lattice points located in the blind region cannot be brought to the diffraction condition using the rotation method around a single axis (Figure 2-7). Nevertheless, this region can be reduced by using a shorter wavelength.

Another factor that should be carefully considered is the exposure time which influences the reflection intensities. Longer exposure time increases the signal-to-noise ratio and therefore improves the data quality but the crystal could suffer from radiation damage due to prolonged X-ray exposure and potentially ruin the later images. In the beam time available, it is a priority to maximally collect a complete dataset even with fairly low intensity rather than an incomplete data set of the better quality.

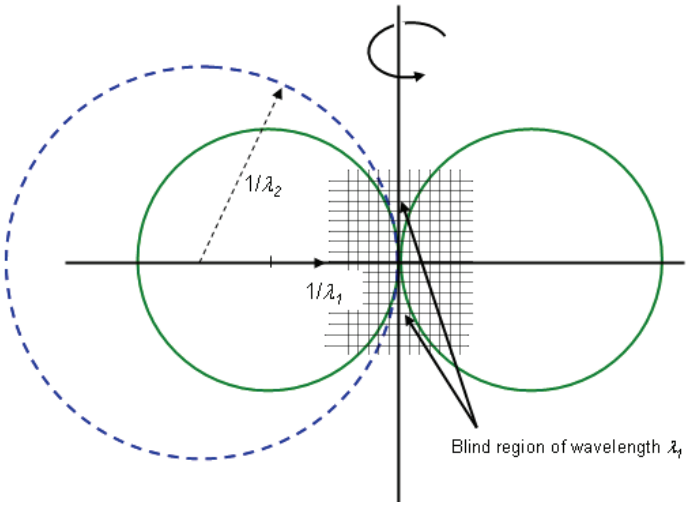


Figure 2-7 Graphical representation of rotation method
Rotating the Ewald sphere around the crystal rational axis will never bring the reciprocal lattice points located in the blind zone to the diffracting condition. A shorter wavelength results in a longer sphere radius ($1/\lambda_2$) and therefore narrows down the blind region.

2.6 DATA PROCESSING

The objective of data processing is to identify the Miller indices (*hkl*) of all reflections recorded on the images, the diffraction intensities I_{hkl}, and their standard deviation σI_{hkl}. This process can be divided into 3 major steps: indexing, integration and scaling (Leslie, 2006). The objective of indexing is to determine the indices *h*, *k*, *l* of each reflection, unit cell dimensions and the orientation of the cell edges with respect to the detector and X-ray beam. In addition, the Bravais lattice symmetry and the mosaicity of the crystal will also be estimated. These parameters will be refined using different algorithms. Several segments of images will be chosen to refine against crystal parameters (including unit cell, crystal orientation and mosaicity), detector parameters (detector position and orientation) and beam parameters (orientation of the primary beam and beam divergence) (Leslie, 2006). After these parameters refinements, all images can be integrated while post-refining the rest of the parameters.

The integration step is to predict the diffraction spots in all images and estimate the intensity and standard uncertainties for each reflection. MOSFLM performs this using summation integration and profile fitting methods (Leslie, 2006). The summation integration works by defining the background and peak mask for each diffraction spot. To obtain the intensity, the pixel values of the peak region are summed and the sum of the background values are subtracted. This is particularly accurate if the

background value is very low compared to the intensity of the spot provided the spots are well resolved. For weak diffracted spots, profile fitting procedures can better estimate the intensity (Leslie, 2006).

The final step in data processing is data reduction or scaling which can be carried out by the CCP4 programme SCALA. The main task of scaling is to bring all reflection intensities to a common scale by modelling the data collection experiment using the intensity values of symmetry equivalent and multiple recorded reflections collected at different times during the experiment (Evans, 2006). Different factors such as radiation damage, X-ray absorption and uneven crystal rotations are modelled and correction factors are calculated (Otwinowski *et al.*, 2003). Two geometric factors are corrected in the scaling process. The first one is the Lorentz factors which account for the fact that the diffracting planes which satisfy the diffracting condition do not occupy such positions for an equal amount of time (Buerger, 1940). The second geometric factor is the polarisation factor which accounts for the fact that electrons do not scatter along their direction of vibration but in other directions with an intensity proportional to $(\sin \alpha)^2$ (Drenth, 2007). These corrections are normally applied during data integration and the programme that integrates the reflections produce Lorentz and polarisation corrected intensities. All the correction factors are optimised in a way that minimises the intensity differences of symmetry related reflections.

At the end of data reduction, various parameters are calculated to give an overview of the quality of all reflections. The log file produced contains useful information such as actual resolution, bad regions of data which should be excluded, anomalous signals for MAD and SAD phasing and overall quality of the dataset (Evans, 2006). The internal consistency of the dataset can be measured as an *R* factor. Examples of these are the R_{merge} or R_{sym} value, which can be expressed as:

$$R_{merge} = R_{sym} = \frac{\sum_{hkl}\sum_{i}\left|I_{hkl,i} - \overline{I_{hkl}}\right|}{\sum_{hkl}\sum_{i}\overline{I_{hkl}}} \qquad \textbf{Equation 2-2}$$

where I_{hkl},i is the ith measurement the reflection *hkl* and $\overline{I_{hkl}}$ is the average intensity of the i observations. The value of this parameter expresses the agreement between symmetry related reflections. R_{merge} tends to increase with the resolution and

multiplicity. Any significant deviation from this value may indicate a problem during data collection. The multiplicity dependent R factor is R_{pim} which can be expressed as:

$$R_{pim} = \frac{\sum_{hkl} \frac{1}{(N-1)^{1/2}} \sum_i \left| I_{hkl,i} - \overline{I_{hkl}} \right|}{\sum_{hkl} \sum_i I_{hkl}} \qquad \textbf{Equation 2-3}$$

where N is the multiplicity of the data. R_{pim} has the advantage over R_{sym} because it is multiplicity dependent.

2.7 ATOMIC SCATTERING FACTOR

The atomic scattering factor or form factor (f_{hkl}) is the sum of scattering from the electron cloud of an atom. It depends on the number of electrons and the scattered angle. The contribution of scattering single atom j to reflection hkl can be written as:

$$f_{hkl} = f_j e^{2\pi i(hxj+kyj+lzj)} \qquad \textbf{Equation 2-4}$$

The atom is treated as a simple sphere of electron density. The function is different for each element depending on the number of electrons (the Z number) which diffract the X-rays. When the scattering angle is zero, f_{hkl} value equals the number of electrons in the atom and this value decreases when the scattering angle increases.

2.8 THE STRUCTURE FACTORS

The structure factor $\mathbf{F_{hkl}}$, the expression for the particular Bragg reflection hkl, is the sum of atomic scattering factors (f_{hkl}) for all atoms in the unit cell. For a given hkl plane, the structure factor can be written as the summation:

$$\mathbf{F_{hkl}} = \sum_j^n f_j e^{2\pi i(hxj+kyj+lzj)} \qquad \textbf{Equation 2-5}$$

Each atom contributes to the structure factor $\mathbf{F_{hkl}}$ and depends on two factors: (i) type of element f_j which contributes the amplitude and, (ii) atom j position in the unit cell (x_j, y_j, z_j). The total structure factor for a lattice point contains both the amplitude $|\mathbf{F_{hkl}}|$ and the phases (α) which can be written as:

$$\mathbf{F_{hkl}} = \left|\mathbf{F_{hkl}}\right| e^{i\alpha_{hkl}} \qquad \textbf{Equation 2-6}$$

The amplitude $|\mathbf{F_{hkl}}|$ is proportional to the square root of the measured intensity (I_{hkl}) of diffraction and phase (α_{hkl}) which is $2\pi(hx_j+ky_j+lz_j)$ in Equation 2-5. The intensity of diffraction depends on atomic scattering factors and position of the atoms (x_j, y_j, z_j) in the unit cell. During data collection, only the intensity of diffraction (I_{hkl}) is

measured by the detector but not the phases and thus only the amplitude of the structure factor can be obtained from the experiment. Therefore, the phase problem needs to be solved in order to determine the arrangement of all atoms in the unit cell.

2.8.1 Friedel's Law

Friedel's law states that the Friedel's pair ($\mathbf{F_{hkl}}$ and $\mathbf{F_{\bar{h}\bar{k}\bar{l}}}$) have the same structure amplitude ($|\mathbf{F_{hkl}}| = |\mathbf{F_{\bar{h}\bar{k}\bar{l}}}|$) but the phases have opposite sign (+α and –α). These reflections with equal amplitudes will produce diffraction spots with the same intensity. Friedel's law holds as long as the anomalous scattering is absent for a particular wavelength used in the experiment.

2.8.2 Electron density

Because $\mathbf{F_{hkl}}$ and $\rho(x, y, z)$ can be Fourier transformed reversibly, the electron density can be expressed as follows:

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} \mathbf{F_{hkl}}\, e^{-2\pi i(hx+ky+lz)} \qquad \textbf{Equation 2-7}$$

By substituting $\mathbf{F_{hkl}}$ with $|\mathbf{F_{hkl}}|e^{i\alpha}$, Equation 2-7 can be written as a Fourier summation:

$$\rho(x,y,z) = \frac{1}{V}\sum_{hkl} \left|\mathbf{F_{hkl}}\right| e^{-2\pi i(hx+ky+lz-\alpha_{hkl})} \qquad \textbf{Equation 2-8}$$

The equation indicates that the electron density $\rho(x, y, z)$ can be calculated at every position (x, y, z) in the unit cell. Although the $|\mathbf{F_{hkl}}|$ can be derived from I_{hkl}, the phase angle α_{hkl} cannot be derived from the diffraction pattern. Therefore, an estimation of phase angle must be done in order to calculate the initial electron density map.

2.9 CALCULATING THE PHASES

During data collection, the crystallographer measures the intensities of waves scattered from planes *hkl* in the crystal. The amplitude of the wave $|\mathbf{F_{hkl}}|$ is proportional to the square root of I_{hkl} measured on the detector. All information of the phases (α_{hkl}) is lost. Since there is no direct relationship between the amplitude and their phases, it is impossible to calculate an electron density map without initial phase information (Equation 2.8). The only relationship between amplitudes and phases is via molecular structure or electron density of the diffracting materials (Taylor, 2003).

Therefore, prior knowledge of electron density calculated from heavy atom scattering or known molecular structure can lead to initial phases of macromolecule. Table 2-2 summarises various phasing techniques along with their prior knowledge that is required to obtain the phases for an unknown structure.

Table 2-2 Phasing techniques

Phasing methods	Prior structural knowledge / assumptions
Direct methods	The electron density is positive or equal to zero, atoms can be described as discrete sphere
Molecular replacement	A homologous structure is available for the unknown structure
Isomorphous replacement	The position of heavy atoms attached to the macromolecule can be determined
Anomalous scattering	Anomalous scatterers attached to the macromolecule can be located

2.9.1 Patterson functions

The Patterson function is a function to calculate electron density from reflection intensities without the use of phase information (Patterson, 1934). This function is a Fourier summation with intensities as coefficients which can be written as:

$$\rho(u,v,w) = \frac{1}{V}\sum_{hkl}\left|\mathbf{F_{hkl}}\right|^2 e^{-2\pi(hu+kv+lw)} \qquad \textbf{Equation 2-9}$$

Where u, v and w are relative coordinates in the Patterson cell that has dimensions equal to the real cell. In terms of convolution, the Patterson function $\rho(u)$ can also be written as:

$$\rho(u,v,w) = \int_{xyz}\rho(x,y,z)\times\rho(x+u,y+v,z+w)dxdydz \qquad \textbf{Equation 2-10}$$

Equation 2-10 illustrates the physical interpretation of Patterson function. This expression indicates that electron density at point (u,v,w) in a Patterson function is a convolution resulting from multiplying the electron density at all points x, y, z by the electron density at points $x + u$, $y + v$ and $z + w$. This will generate peaks corresponding to the atomic numbers of the two atoms at the ends of the vector u, v and w in the map. The Patterson map is then an inter-atomic vector map and contains information for the distances between atoms. If a real unit cell contains N atoms, the corresponding Patterson map will show N^2 peaks. Of these, N peaks are superposed as

a result of self-convolution, leaving non-origin peaks of N^2-N. Additionally, the Patterson map is centrosymmetric at the origin. For each pair of atoms there will be two Patterson peaks due to the two vectors with equal magnitude but in opposite directions. The Patterson method can be used to solve small molecule structures with a limited number of atoms but this does not apply to protein structures which typically contain thousands of non-hydrogen atoms. Nevertheless, the Patterson function is particularly useful to locate a limited number of heavy atoms in a large unit cell since the height of peaks in the Patterson map depends on atomic number.

2.9.2 Experimental phasing

When a novel X-ray structure cannot be solved by molecular replacement because a homologous model is unavailable, experimental phasing must be used to solve the structure crystallographically. Experimental phasing methods include single and multiple isomorphous replacements (SIR and MIR), single and multiple isomorphous replacement with anomalous scattering (SIRAS and MIRAS) and single and multiple wavelength anomalous dispersion (SAD and MAD). All these methods rely on the presence of heavy atoms in the macromolecular crystal in a high occupancy to give a clear signal.

The first step in determining experimental phases is introducing heavy atom into the crystal to produce measurable changes in intensities. Protein crystals are open lattices that contain a high percentage of solvent. These solvent channels allow heavy atom compounds to diffuse into the crystal by soaking. Screening for an adequate compound for a given protein crystal usually involves trial and error before a useful derivative is found. Heavy metals can bind to protein in a number of distinct ways. For example, single metal ions are able to interact with the protein surface electrostatically; mercurial compounds tend to bind Cys sulphur or His nitrogen and platinum compounds bind to Cys, His and Met (Garman and Murray, 2003). The incorporation of heavy atoms such as replacing Met with seleno-Met by a genetic approach become a robust method (Hendrickson *et al.*, 1985). To solve a novel structure of a nucleic acid, it is convenient to introduce heavy atoms such as 5'iodo- or 5'bromo-uracil into oligonucleotides by direct incorporation during the nucleic acid synthesis.

Nonetheless, several problems could adversely affect experimental phasing. One of the biggest problems of heavy atom derivatisation is non-isomorphism which induces a change in unit cell dimensions and disturbs the molecular arrangement in the crystal. This problem entirely jeopardises experiments such as SIR, MIR, SIRAS, and MIRAS. A common obstacle in heavy atom soaking is that the macromolecular crystal could dissolve in the heavy atom compound solution during soaking. Oxidation of selenium in seleno-Met derivatives causes the *K*-edge to move to a higher energy and the "white line" to be enhanced owing to changes in the chemical environment of the selenium during oxidation (Garman and Murray, 2003). Radiation damage is the common problem in MAD experiments due to the strong absorption properties of heavy atoms.

2.9.3 Isomorphous replacement

In the isomorphous replacement methods (SIR and MIR), the X-ray diffraction pattern of the native structure is compared to that of the heavy atom derivative crystal. Ideally, the conformation of the protein, unit cell dimensions, as well as orientation and position of the macromolecules in the native and in the derivative crystal are exactly the same. In this way, the intensity differences between the native and derivative crystal are exclusively due to the presence of heavy atoms. From these differences, the heavy atom position can be determined using Patterson function [eg. SIR2002 (Burla *et al.*, 2004)], direct methods [eg. SHELXD (Schneider and Sheldrick, 2002) and SnB (Weeks and Miller, 1999)].

The structure factors from the unmodified protein crystal are the sum of the scattering vectors from the protein atoms. Whereas the structure factors from a heavy atom derivative are the sum of the scattering vectors from all the protein atoms plus the heavy atom. The isomorphous replacement method assumes that the scattering vector from the protein atoms is the same in both the native and derivative crystals. Therefore, the relationship of the native and derivative structure factors is:

$$\left|\mathbf{F_H}\right| = \left|\mathbf{F_{HP}}\right| - \left|\mathbf{F_P}\right| \qquad \textbf{Equation 2-11}$$

where $\mathbf{F_H}$ is the structure factor of heavy atoms; $\mathbf{F_{HP}}$ is the structure factor of protein atoms plus heavy atoms and $\mathbf{F_P}$ is the structure factor of protein atoms. To calculate

the $\mathbf{F_H}$, replacing Equation 2-11 in the Patterson function equation (Equation 2-9), will give the isomorphous difference Patterson function as below:

$$\rho(u,v,w) = \frac{1}{V}\sum_{hkl}\left(|\mathbf{F_{HP}}| - |\mathbf{F_P}|\right)^2 e^{-2\pi i(hu+kv+lw)} \quad \textbf{Equation 2-12}$$

The heavy atom position in the unit cell can be determined from an isomorphous difference Patterson map. Together with the amplitudes $|\mathbf{F_{HP}}|$ and $|\mathbf{F_P}|$, the phases of $\mathbf{F_P}$ can be estimated using a Harker construction (Figure 2-8).

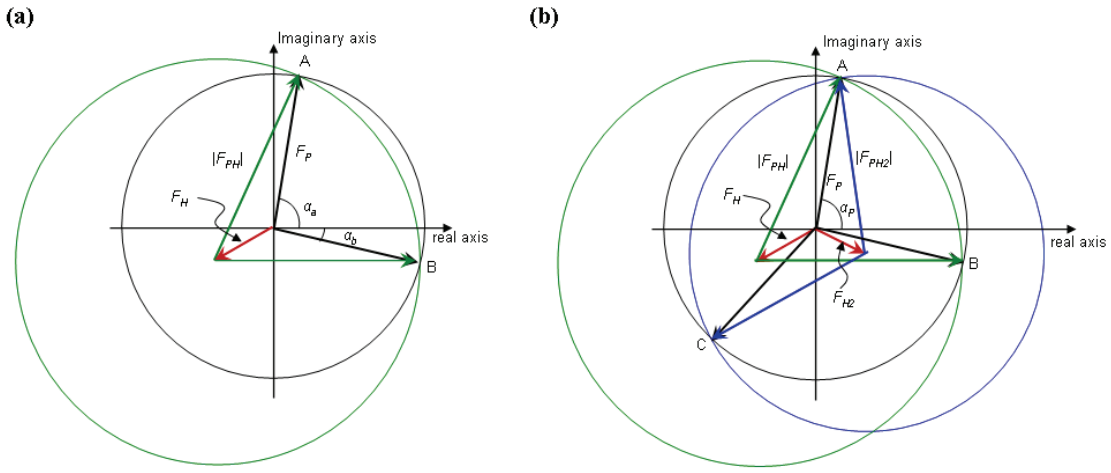


Figure 2-8 Harker constructions of SIR and MIR
(a) Phase calculation of single heavy atom in SIR and (b) two heavy atoms in MIR. (Diagram adapted from Taylor 2003).

In SIR (Figure 2-8a), all the possible phases of $\mathbf{F_P}$ and $\mathbf{F_{PH}}$ are shown as black and green circles, respectively. The radius of the circles are represented by $|\mathbf{F_P}|$ and $|\mathbf{F_{PH}}|$ and the intercept points A and B represent the possible correct phases for $\mathbf{F_P}$ (α_a and α_b); in which; both agree with the measured $|\mathbf{F_{PH}}|$ and calculated $\mathbf{F_H}$. This phase ambiguity can be resolved by MIR. The $|\mathbf{F_{PH}}|$ of a second derivative is measured and resulted a third interception (point C) as shown in the Harker construction for MIR (Figure 2-8b). Only the vector to the intercept point (point A) which agreed with all measured and calculated **F**s give the correct phase angle for $\mathbf{F_P}$.

Another way to solve the ambiguity in phase is using the heavy atoms which also serve as anomalous scattering atoms in isomorphous replacement. Where the Bijvoet difference [$\mathbf{F_{PH}}$(+) and $\mathbf{F_{PH}}$(-)] can be measured reliably, anomalous scatterers may function as the second heavy atom derivative, as in an MIR experiment. The complementary use of anomalous scattering and single isomorphous derivative data can give unique phases. This method is called single isomorphous replacement with

anomalous scattering (SIRAS). On the other hand, multiple isomorphous replacement with anomalous scattering (MIRAS) can be used with several heavy atom derivatives which give anomalous dispersion signals, and the Bijvoet pairs can be measured for each. The additional intensity observation will improve the accuracy of phase determination.

2.9.4 Anomalous scattering

The X-ray absorption of an element is a function of the X-ray wavelength and a sudden drop in absorption at a specific wavelength is referred to as the adsorption edge of the element. At the adsorption edge, an electron is ejected from an atom by the photon energy of the X-ray beam. The vacated position is filled by an electron drop from a higher shell (eg. *L* shell to *K* shell of copper) and this transition causes the element to emit at its characteristic wavelength. This emitted ray carries a different energy and the element is said to exhibit anomalous scattering at the absorption edge. The absorption edge of light atoms for instance oxygen, carbon and nitrogen are not close to the wavelength of the X-rays used in crystallography. As a result, the light atoms do not contribute to anomalous scattering. However, absorption edges of heavy atoms (such as selenium and manganese) are within this range. Anomalous scattering causes the anomalous scattered ray to have different intensities and phases.

The anomalous scattering causes both the intensity and phase of the anomalous scattered rays to be altered from the normal atomic scattering factor f_o. Therefore, the total scattering factor for anomalous scattering f can be written as:

$$f = f' + i\Delta f''$$
$$f = f_o + \Delta f' + i\Delta f''$$

Equation 2-13

Where f' and $\Delta f''$ (usually referred to as f'') are the real and imaginary parts, respectively, of the anomalous scattering factor. $\Delta f'$ and $\Delta f''$ are also known as the dispersion and absorption components of the anomalous scattering, respectively. The $\Delta f'$ has the same phase as f_o whereas $\Delta f''$ has a phase $\pi/2$ with respect to f_o.

The theoretical values of f' and f'' (Cromer and Liberman, 1970) can be computed and displayed using web interface such as http://skuld.bmsc.washington.edu/scatter. However, the extreme positions on the curves are affected by the chemical

environment of the heavy atom. Therefore, the values of *f'* and *f''* must be determined using an X-ray fluorescence spectrum for the protein crystal under investigation. Figure 2-9 shows a theoretical anomalous scattering of selenium around its absorption edge.

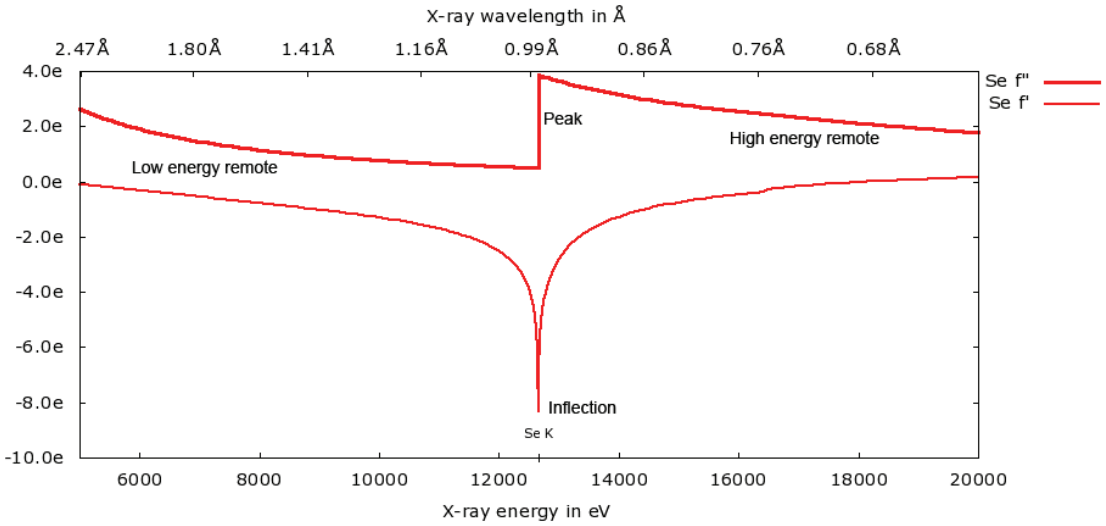


Figure 2-9 Anomalous scattering around selenium absorption *K*-edge
The *f''* and *f'* values are shown at the maximum and minimum values of 'peak' and 'inflection', respectively. The regions far from the absorption edge are corresponding to high and low remote energies. (Diagram was generated from http://skuld.bmsc.washington.edu/scatter/AS_form.html).

2.9.5 Breakdown of Friedel's Law

If atoms H exhibiting anomalous scattering coexist with the non-anomalous scattering atoms P, the structure factor of the Friedel's pair can be written as:

$$\mathbf{F}_{\mathbf{PH}}(+) = \mathbf{F}_{\mathbf{P}}(+) + \mathbf{F}_{\mathbf{H}}'(+) + i\,\mathbf{F}_{\mathbf{H}}''(+)$$
$$\mathbf{F}_{\mathbf{PH}}(-) = \mathbf{F}_{\mathbf{P}}(-) + \mathbf{F}_{\mathbf{H}}'(-) + i\,\mathbf{F}_{\mathbf{H}}''(-) \qquad \textbf{Equation 2-14}$$

Where, $\mathbf{F_P}(+)$ and $\mathbf{F_P}(-)$ are normal protein structure factors without anomalous scatter, $\mathbf{F_H}'(+)$ and $\mathbf{F_H}'(-)$ are real part of the anomalous structure factor (dispersive components) and $i\mathbf{F_H}''(+)$ and $i\mathbf{F_H}''(-)$ are the imaginary part of the anomalous structure factor (absorption components).

Their contribution to the vector PH for each Friedel's pair $\mathbf{F_{PH}}(+)$ and $\mathbf{F_{PH}}(-)$ can be represented as Figure 2-10. As a result, the intensities of a reflection *hkl* and its Friedel's pair $\overline{h}\overline{k}\overline{l}$ are no longer equal and therefore Friedel's law is breached. The difference; $\mathbf{F_{PH}}(+)^2 - \mathbf{F_{PH}}(-)^2$, is known as the Bijvoet difference or the anomalous

scattering differences. These differences are useful in finding the anomalous scattering atoms in the crystal.

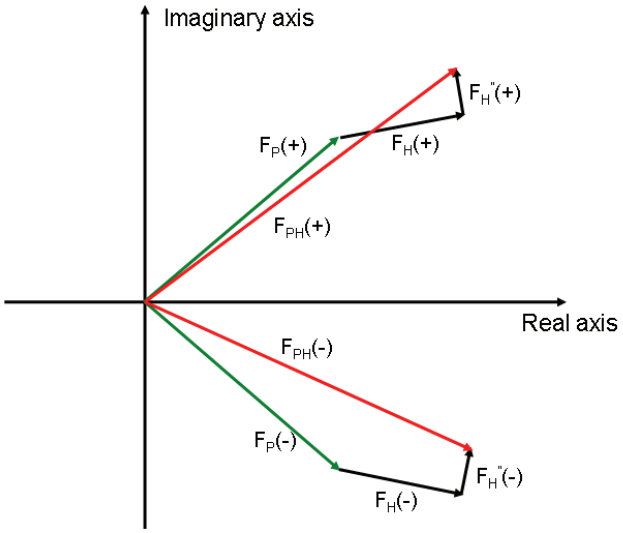


Figure 2-10 Representation of the vector summation of the anomalous scattering
Argand diagram representing the vector summation of the anomalous scattering contribution of atoms H to normal scattering of atoms P (Adapted from Blundell, 1976).

2.9.6 Single- and multi-wavelength anomalous dispersion (SAD and MAD)

If data sets at different wavelengths can be collected from a single crystal containing heavy atoms that give anomalous dispersion, then a MAD phasing experiment can be carried out to solve the structure. Tuneable synchrotron wavelengths enable the measurement to be made precisely at the chosen wavelengths. The advantage of MAD over SIR/MIR is that the non-isomorphism problem would not occur because the measurements of native and derivative structure factors are within one single crystal. For proteins that naturally contain a heavy atom such as metalloproteins, these heavy atoms can provide a source of anomalous dispersion. For proteins that lack heavy atoms, sulphur-Met can be replaced with seleno-Met in protein expression (Hendrickson *et al.*, 1990). Moreover, the introduction of methionine into the protein by site-directed mutagenesis has been shown to allow phasing from a selenomethionyl-substituted protein crystal (Leahy *et al.*, 1994). In general, the presence of one selenium atom in a protein containing less than 150 residues is sufficient to perform MAD (Hendrickson *et al.*, 1990).

For MAD phasing, it is usual to collect three datasets at different wavelengths, known as 'peak' (λ1), 'edge' (inflection) (λ2) and 'remote' (λ3) around the absorption edge. At the 'peak' wavelength, the imaginary part *f"* should have largest value; at 'edge' wavelength, the real part of *f'* should have its lowest value and at 'remote' wavelength, it is either longer (high remote) or shorter (low remote) than 'peak'/'remote', where *f'* and *f"* are small (Figure 2-9). Wavelengths λ1 and λ2 provide data rather similar to two isomorphous derivatives and λ3 functions as a parent. MAD can be applied as a special case for MIR by exploiting the anomalous scattering of existing atom. The first step in MAD is locating heavy atom position in the unit cell using anomalous difference Patterson function by coefficient:

$$\Delta\left|\mathbf{F}_{\mathbf{ano}}\right|^2 = \left(\left|\mathbf{F}_{\mathbf{PH}}(+)\right| - \left|\mathbf{F}_{\mathbf{PH}}(-)\right|\right)^2 \qquad \textbf{Equation 2-15}$$

Where $\left(\left|\mathbf{F}_{\mathbf{PH}}(+)\right| - \left|\mathbf{F}_{\mathbf{PH}}(-)\right|\right)^2$ represents Bijvoet difference amplitudes. The heavy atom vector map can then be generated to locate the heavy atom position. Accurate intensity measurement is important because the anomalous signal is relatively small compared to an isomorphous difference Patterson.

Disadvantages of MAD are that multiple wavelengths have to be observed which increase the X-ray exposure time and the danger of radiation damage to the crystal. These disadvantages can be minimised if the crystal structure can be solved with a single wavelength. In SIR, reflection data from a single wavelength is used to provide strong phase information but both native and derivative crystals are required. SAD is an alternate method to solve X-ray structures containing an anomalous scatterer with a single dateset at an appropriate wavelength. In principle, any anomalous scatterer can be used for SAD. As for MAD, the first step in applying SAD is to locate the anomalous scatterers from an anomalous Patterson map using coefficients $\left|\left|\mathbf{F}(hkl)\right| - \left|\mathbf{F}(\bar{h}\bar{k}\bar{l})\right|\right|^2$ or by direct methods as in SHELXD (Schneider and Sheldrick, 2002; Uson *et al.*, 2003) or SnB (Hauptman, 1997a; Hauptman, 1997b). This information is indicated in the Argand diagram shown in Figure 2-11a. The phase ambiguities (Figure 2-11b) generated from SAD is similar to SIR. In order to determine the correct enantiomorph, it is usual to refine both substructures and select the better of the two resulting electron density maps.

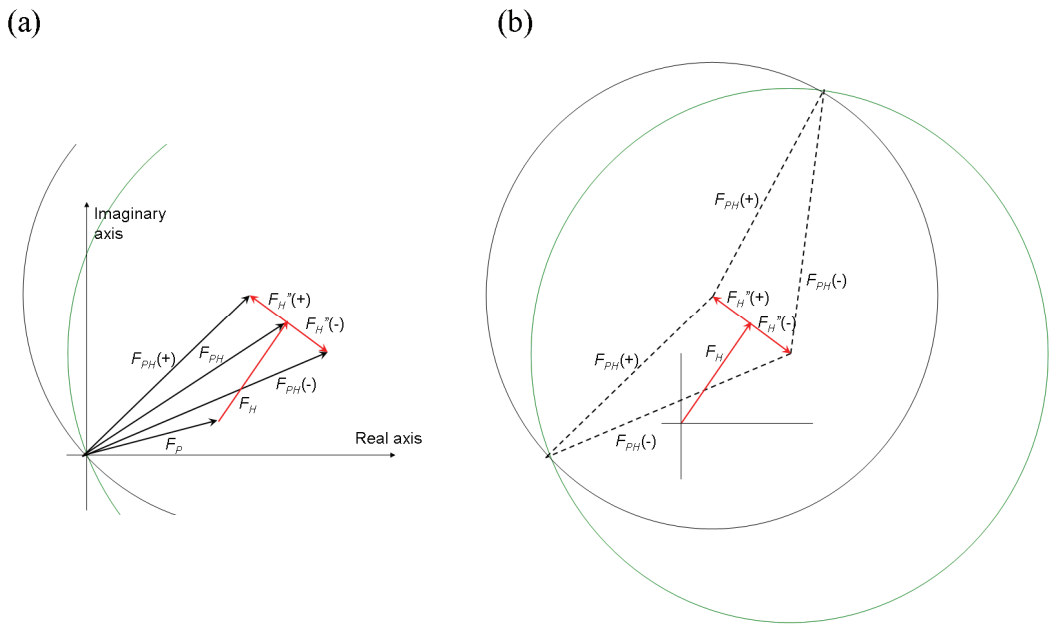


Figure 2-11 Illustration of SAD phasing
(a) Argand diagram of SAD and (b) Harker construction for single wavelength SAD. The red lines in opposite directions indicate two possible phase angles as a result of $\mathbf{F_H}$"(+) and $\mathbf{F_H}$"(-) (diagrams modified from Drenth 2007)

2.9.7 Molecular replacement

Molecular replacement is a technique to solve the phase problem using a known structural model; namely phasing model; to calculate initial phases of an unknown structure. To successfully solve a new structure, the phasing model usually has a high percentage of amino acid sequence similarity (usually above 25%) and a similar polypeptide folding with the unknown protein. The main task of molecular replacement is to place the homologous model into a proper orientation and position in the target unit cell by 6 parameters (three rotational and three translational functions). Searching all 6 parameters simultaneously is computationally expensive. Luckily, some properties of the Patterson function can be used to define these parameters. The vectors of the Patterson map can be divided into two; intramolecular vectors (vectors between atom in the same molecule) and inter-molecular vectors (vectors between atoms in two molecules). Intramolecular vectors depend on the orientation of the molecule, not its position in the unit cell. These vectors are shorter than the inter-molecular vectors and are confined to a region near the Patterson map origin. When the molecule is rotated, these vectors rotate in the same direction as the molecule. The three parameters in describing orientation are the Euler angles (α, β, γ).

The model Patterson is rotated by an angle α around the original axis z, then by an angle β around axis y and followed by an angle γ around z axis, and then superimposed on the observed Patterson. A new orientation of the model Patterson will be generated.

After the orientation of the phasing model has been determined by the rotation function, the next step is to place the model in the correct position by the translation function. The translation function can be performed using a cross-Patterson function. Cross-Patterson vectors in the Patterson map are derived from inter molecular vectors resulting from symmetry related models. A translation function is a calculation of correlation between cross-Patterson vectors of the model structure and the observed Patterson function.

After the translation function, the structure factors of the phasing model ($\mathbf{F_{calc}}$) are calculated and compared with the observed structure factors ($\mathbf{F_{obs}}$) by calculating an *R*-factor (Equation 2-16). The agreement between $\mathbf{F_{obs}}$ and $\mathbf{F_{calc}}$ can also be compared using a correlation coefficient which is insensitive to the scaling factor. The phases of the unknown structure are estimated from the phase angles (α_{hkl}) implied by the proper oriented and positioned phasing model which must be attached to the $|\mathbf{F_{obs}}|$ in order to calculate an electron density map.

2.10 DENSITY MODIFICATION AND PHASE IMPROVEMENT

The initial set of phase angles obtained from molecular replacement or experimental phasing contains errors and electron density maps calculated from these phases are often not readily interpretable. Density modification procedures aim to improve the electron density map and the phases by imposing some physical constraints on the electron density. A new set of amplitudes and phase angles will be calculated and therefore improve the electron density maps. The new amplitudes together with the experimental amplitudes are used to calculate appropriate weights for the combination of the new set of phases with the initial phases. The process is recycled until convergence. Three major techniques used in density modification are solvent flattening, histogram matching and non-crystallographic symmetry averaging [reviewed in Cowtan and Zhang (1999)].

Protein crystals contain a high percentage of solvent which contribute to a poor diffraction pattern because solvent components are not ordered. In the solvent flattening method, the electron density is divided into solvent and protein regions (Wang, 1985). The solvent region is modified and set to a mean value while the protein region of the map does not alter. The new electron density distribution is used to calculate structure factors and the phases of which are combined with the experimental phases. The combined phases and the observed structure factors are used to calculate an improved electron density map. The resulting map will look much flatter in the region assigned as 'solvent region'. The process may be repeated cyclically until convergence. A complementary method to solvent flattening is histogram matching which may be applied to the protein region of the map (Zhang and Main, 1990). Given the resolution, overall B-factor, overall scale and other parameters, the expected form of histogram of electron density values can be predicted with precision. This method attempts to reassign all region of electron density within the protein volume to a value consistent with the predicted distribution of densities for the structure at the observed resolution. The modified electron density map can be used to calculate a new set of phase angles with improved accuracy, from which, an improved electron density map can be produced. The procedure is repeated until convergence.

If there is more than one identical subunit in the asymmetric unit which is related by non-crystallographic symmetry (NCS), the electron density can be modified by NCS symmetry averaging. By knowing the type of NCS elements and the location of the molecules, the subunits can be moved to an identical orientation by a rotation function (namely self-rotation function) and the target and model Pattersons are summed and averaged. The averaged map displays a clearer image with higher signal-to-noise ratio. Algorithms to perform these three density modification techniques are implemented in DM (Cowtan, 1994) and RESOLVE (Terwilliger, 2003c)

2.11 MODEL BUILDING

The quality of electron density map resulting from phasing and density modification determines the model building strategies. Model building is the process of fitting a

realistic molecular model into the electron density map. For a good quality map (usually better than 2.5 Å) with clear protein main-chain connectivity and visible amino acid side-chains, the model can be built automatically using programmes such as ARP/wARP (Morris *et al.*, 2003; Morris *et al.*, 2004) and RESOLVE (Terwilliger, 2003c). In brief, the strategy used in ARP/wARP involves identifying the peaks as free atoms (free-atom models) and the quality of phases is improved based on averaging techniques using the atomic coordinates. Protein elements can then be automatically recognised as peptide planes thus a partial atomic protein model can be built using the main-chain autotracing feature. The hybrid model (containing free atom and partial built model) is then refined using new stereochemical data and this improves the accuracy of the phases. Afterwards, the amino acid side-chain will be identified and sequence docking is performed onto the already built main-chain. After each cycle, new phases will be calculated and recombined with phases from the previous cycle to produce an improved map. These steps are iterated until no further improvement can be achieved.

For poorer quality electron density maps (usually below 2.5 Å), the atomic picking approach in ARP/wARP becomes less efficient. Automatic building however can be still attempted using RESOLVE which claims still to function with data as low as 3 Å in resolution (Terwilliger, 2004; Terwilliger, 2003a; Terwilliger, 2003b). This approach builds the atomic model into the electron density map by linking predefined protein fragments selected from pre-compiled libraries. The building cycles are interspersed with refinement and density modification and the convergence output usually provides a good model for manual corrections using graphical programmes such as COOT (Emsley and Cowtan, 2004).

2.12 REFINEMENT

After a rather complete model is available its parameters have to be optimised in order to give a best possible fit to the crystallographic observations. The experimental data such as unit cell parameters, structure factor amplitudes, and experimental phases are considered as observations. For observations to be used in refinement, the consistency between them can be formally related with the variable model parameters such as atomic coordinates, B-factor, and occupancies through refinement functions like least

squares and maximum likelihood (Tronrud, 2004). In practice, the three positional parameters (x, y, and z) and only one isotropic B factor are refined for each non-hydrogen atom. The number of reflections at a given resolution can be used to calculate the ratio of observations to parameters. For a poor ratio, additional observations such as geometrical restraints are incorporated into the refinement process. Constraints are applied when rigid geometry such as amides are involved and only the dihedral angles need to be varied. These assumptions are usually precise and valid for macromolecular structures (Engh and Huber, 1991). Restraints are applied when the stereochemical parameters such as bond lengths, bond angles, torsion angles and van der Waals contacts are allowed to vary around a standard value but penalties are applied when calculated values vary from target values (Drenth, 2007).

2.13 MODEL VALIDATION

The progress of model building and refinement should be monitored in a way that the observed and model structure factors converge in the final stages of structural refinement. There are several ways to inspect the accuracy of the final model. The conventional *R* factor (R_{cryst}) compares the observed structure amplitudes $|\mathbf{F_{obs}}|$ to those calculated from the model $|\mathbf{F_{calc}}|$ can be defined as:

$$R_{cryst} = \frac{\sum_{hkl} \left| |\mathbf{F_{obs}}| - k|\mathbf{F_{calc}}| \right|}{\sum |\mathbf{F_{obs}}|} \qquad \textbf{Equation 2-16}$$

The refinement is carried in the working set only and the free R-factor (R_{free}) is calculated with the test set of reflection only:

$$R_{free} = \frac{\sum_{hkl} \left| |\mathbf{F_{obs}}| - k|\mathbf{F_{calc}}| \right|}{\sum_{hkl} |\mathbf{F_{obs}}|} \qquad \textbf{Equation 2-17}$$

Both *R*-factors should gradually decrease when $\mathbf{F_{calc}}$ approaches $\mathbf{F_{obs}}$. Stereochemically, the protein main-chain can be specified using two conformational angles, φ and ψ, for each C_α atom. All these angles can be displayed using a Ramachandran plot (Ramachandran *et al.*, 1963). The final validation of the model can be carried out using programmes such PROCHECK (Laskowski *et al.*, 1993), SFCHECK (Vaguine *et al.*, 1999) and COOT (Emsley and Cowtan, 2004).

CHAPTER 3. PROTEIN PURIFICATION AND CHARACTERISATION

3.1 INTRODUCTION

The first step of macromolecular crystallisation is to obtain a homogenous and biophysically active macromolecule. The quantity and quality of mammalian protein required for crystallisation is difficult to achieve from mammalian cell culture system due to low expression profile and heterogeneity of mammalian cell extract. To highly express mammalian proteins, the coding gene can be incorporated into an *E. coli* expression vector. The gene expression is usually controlled by a T7 promoter located upstream of the RNA transcription site and the protein expression can be induced by addition of isopropyl-β-D-thiogalactopyranoside (IPTG). In order to obtain a homogenous sample, which is a prerequisite in growing macromolecular crystals, the protein can be purified with various biophysical (eg. salt precipitation and ultracentrifugation) and chromatographic techniques (eg. gel filtration, ion affinity and ion-exchange chromatography). The principle behind these techniques is to separate unwanted proteins from the protein of interest. Several methods are usually involved in order to obtain a homogenous sample.

All plasmids carrying the human MeCP2 encoding gene were generous gifts from Dr. Robert J. Klose. The method for MeCP2 purification was initially described by Klose and Bird (2004) but has been modified in order to obtain a larger quantity of pure protein for initial crystallisation screening and biophysical characterisations. A combination of 3-step column chromatography has been used to purify MeCP2 constructs produced using the BL21(DE3)pLys expression system. All constructs were C-terminally tagged with a 6xHis. Thus the first column was Ni-NTA/Talon metal affinity binding chromatography, which was followed by Sephacryl S200 gel filtration column, and lastly, Sp-Sepharose cation exchange chromatography. All protein required for co-crystallisation was freshly prepared and stored at 4 ˚C in a relatively high salt buffer (20mM HEPES pH7.6, 300mM NaCl). The sample is stable for 1-2 months. Purified protein intended for biophysical analysis can be stored at -80 ˚C in the presence of 10 % (v/v) glycerol for approximately 1-2 years.

3.1.1 Protein chromatography

3.1.1.1 Immobilised Metal Affinity Chromatography

All MeCP2 constructs were designed to have an uncleavable 6xHis tag at their C-terminus, in which, the translationally premature proteins can be eliminated from co-purification. Therefore, the first column used to separate the protein of interest from crude lysate was Ni-NTA or Talon affinity column. In principle, the Ni^{2+} is chelated onto a nitrilo triacetic acid (NTA) resin and protein sample is passed through a Ni-NTA column (Figure 3-1). Recombinant protein containing the 6xHis-tag binds strongly to the charged resin through interactions between the immobilised metal ions (Ni^{2+} or Co^{2+}) and the polypeptide tails whereas only a few naturally occurring metal binding proteins present in the sample will bind weakly to the charged resin. These weakly bound proteins can then be competitively removed with buffer containing a low concentration of imidazole which mimics the histidine side-chain. Other unspecifically bound proteins can be washed by increasing the imidazole concentration. Lastly, the 6xHis tagged recombinant protein can be eluted with a higher concentration of imidazole.

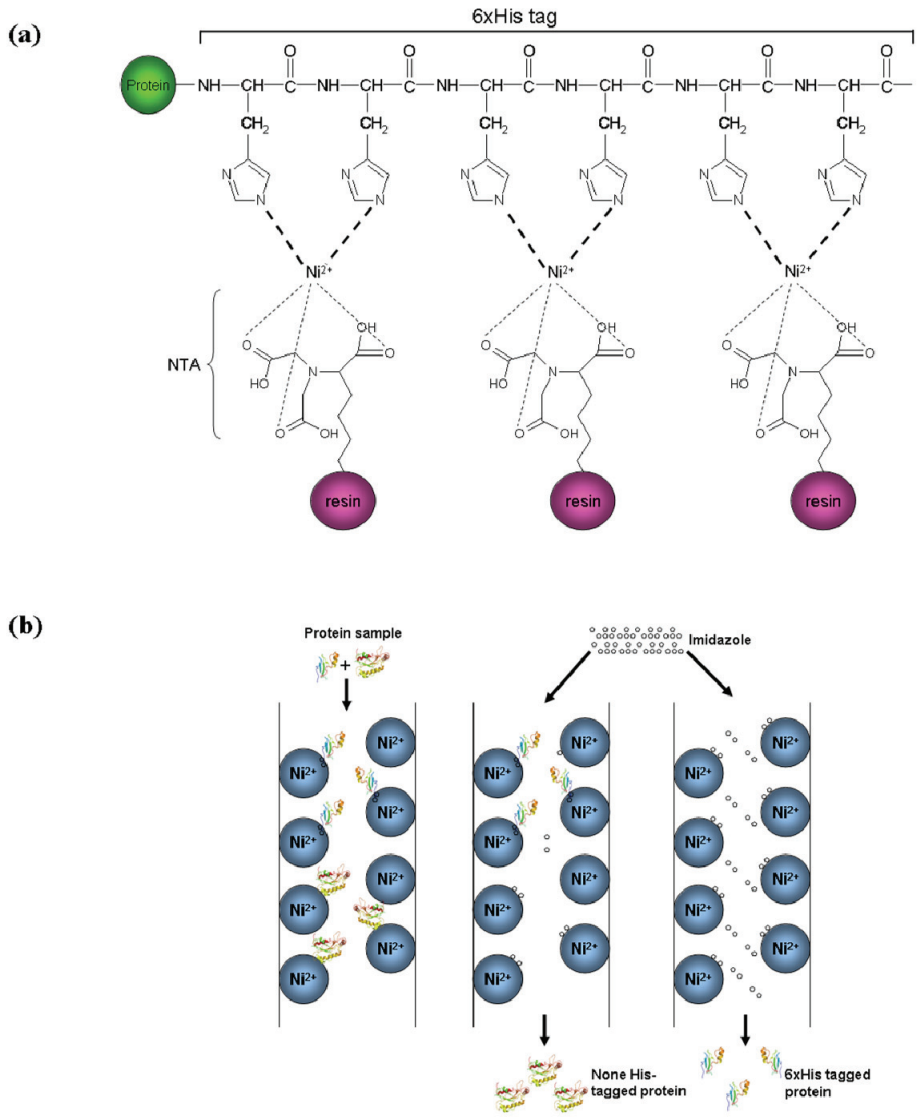


Figure 3-1 Immobilised metal affinity chromatography (IMAC)
(a) Divalent cations such as Ni^{2+} or Co^{2+} are chelated onto a nitrilo triacetic acid (NTA) conjugated resin. Recombinant His-tagged protein binds strongly to the immobilised ions through interactions with the polyhistidine tag. (b) Protein sample is loaded onto the Ni-NTA column, Left panel; proteins with affinity for nickel bind to the column, Middle panel; the presence of low concentration of imidazole removes weak nickel binder and Right panel; strongly bound His-tagged protein is eluted from the column with high imidazole concentration.

3.1.1.2 Sephacryl s200 gel filtration

Since there were considerable contaminants co-eluted using IMAC, the second step was employed to purify MeCP2 constructs was Sephacryl S-200HR gel filtration chromatography. In principle, gel filtration is a purification technique that separates macromolecules based on molecular sizes. The stationary phase of the column contains a porous non-absorbing bead with a well-defined range of pore sizes (Figure 3-2). Small proteins can enter the pores of the stationary phase whereas the larger

proteins are excluded from the pores and only can access into the mobile phase and thus they are eluted first.

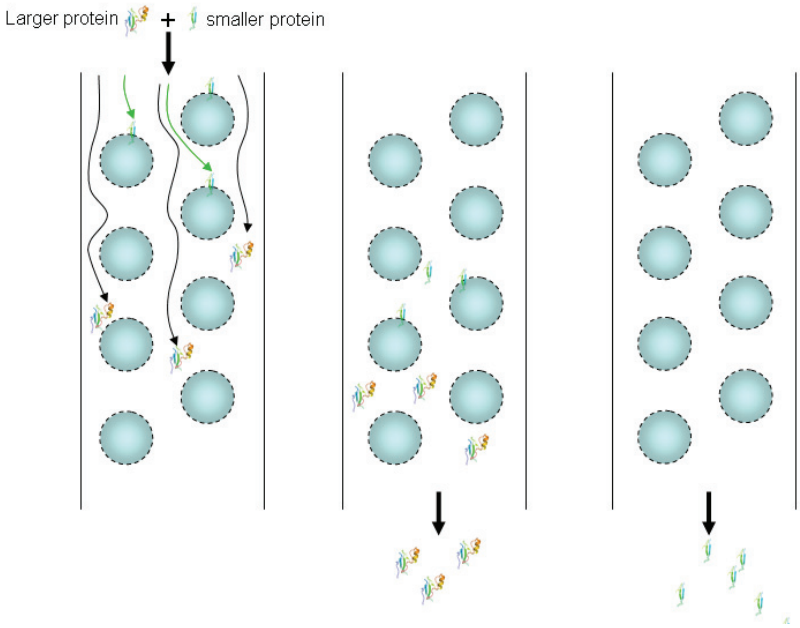


Figure 3-2 Size exclusion chromatography
Protein sample is loaded onto the column containing porous matrix. Left panel shows that higher molecular weight proteins migrate faster than small size proteins which can enter the matrix pores. Middle panel; large proteins are eluted first before (right panel) lower molecular weight proteins.

3.1.1.3 Sp-Sepharose cation exchange chromatography

Ion exchange chromatography is a widely used method for protein purification. The ion exchangers are either positively (anion exchange) or negatively (cation exchange) charged matrices, onto which the charged molecules such as protein can be bound through electrostatic interactions (Figure 3-3). The ionic strength of electrostatic interaction between the charged matrix and the protein is highly dependent on the pH and the ionic strength of the surrounding environment as well as the net charge of the protein. At a low ionic strength environment, all proteins that have affinity for an ion exchanger will bind to the matrix. By increasing the ionic strength of the buffer using a higher salt concentration, salt ions will compete with the proteins for the matrix. Therefore, low affinity bound proteins will be eluted whilst the more tightly binding remains bound. The ionic strength can be increased continuously using a salt gradient onto the column. Proteins with higher affinity for the matrix can be sequentially eluted. An important feature of ion exchange chromatography is that the net charge of the protein can be manipulated by choosing an appropriate pH. Therefore, the binding

strength of the protein onto to the cation or anion exchanger can be changed by altering the buffer pH. As a result, a chromatographic condition can be established for a particular protein by tuning between pH and the ionic strength of a given buffer provided the pI of the protein is known.

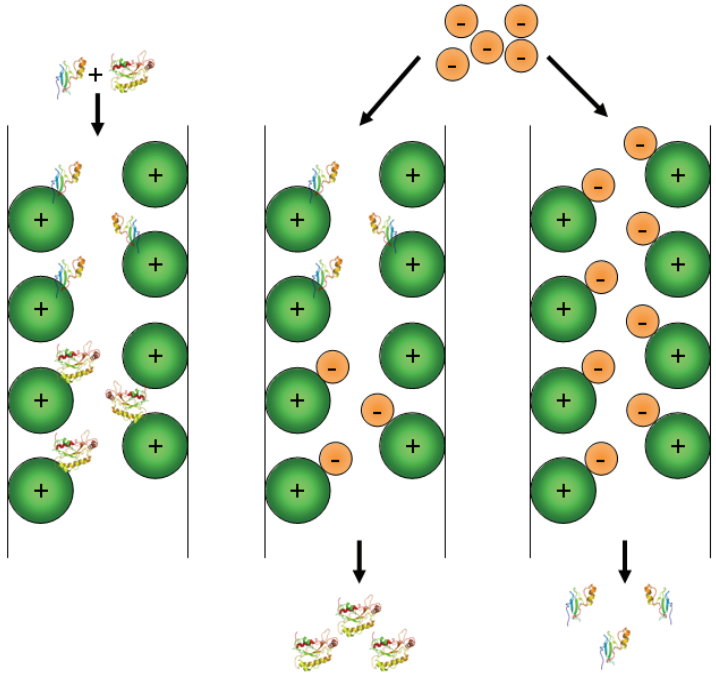


Figure 3-3 Ion exchange chromatography
Protein sample is loaded onto the ion exchange column containing ion exchanger. Left panel; proteins bind to the charged matrix with different affinities depending on their net charge. Middle panel; introduction of a salt gradient gradually alters matrix surface charges and weakly bound proteins can be removed. Right panel; the matrix is fully negatively charged and all proteins are eluted at last.

3.2 MATERIALS AND METHODS

3.2.1 Transformation

DNA plasmid was transferred into a microcentrifuge tube containing competent cells which were defrosted on ice. The transformation reaction was incubated on ice for 5 min. The cells were then heat shocked for 30s at 42ºC on a heat block and immediately followed by a 2 min resting on ice. Room temperature SOC medium [2.5mM KCl, 10mM $MgCl_2$, 10mM $MgSO_4$, 0.4 % (w/v) glucose; 250 µl] was added into the transformation reaction and the mixture was incubated at 37 ºC for 1 h. The transformation reaction was then plated onto LB agar containing an appropriate antibiotic.

3.2.2 Protein expression and purification

The recombinant plasmid encoding MeCP2 was transformed into strain BL21(DE3)pLys (Novagen) as above and plated onto LB agar supplemented with kanamycin (50μg/ml). A single colony was inoculated into 50ml LB medium and incubated at 37˚C overnight. An aliquot of 10ml overnight culture was transferred to 1L LB medium containing kanamycin. Protein expression was induced by addition of IPTG with a final concentration of 1mM when cells grown at 37˚C reached mid-log phase (OD_{600nm} = 0.6). Incubation was continued for an additional 6h at 30˚C before harvesting by centrifugation at 6000*xg* for 20min at 4˚C. Cells were resuspended in 20ml lysis buffer (20mM Tris.HCl pH8.0, 500mM NaCl, 0.1% (v/v) NP40) containing a complete protease-inhibitor mix (Roche) and lysed by sonication using a sonicator (MSE Soniprep 150) at output 15X for a total of 2 min. Crude lysate was recovered by centrifugation at 64,000*xg* for 30min at 4˚C and loaded onto Ni-NTA (Novagen) column on an FPLC system (Applied Biosystems), washed with 12% (w/v) buffer N250 (20mM NaH_2PO_4, 300mM NaCl, 250mM imidazole), and protein was eluted with a gradient from 12-40 (w/v) % with buffer N250. Peaks fractions collected at approximately 25% N250 were shown by SDS-PAGE (18% w/v) to contain MBD protein. Positive fractions were pooled and concentrated to 2ml in a Vivaspin 50ml centrifugal concentrator with a 5 kDa cutoff. The protein was then applied onto a sephacryl-200 gel filtration column on an FPLC system which was pre-equilibrated with GF300 (20mM HEPES pH7.6, 300mM NaCl). Fractions containing MBD protein, as analysed with SDS-PAGE, were pooled, and loaded onto a cation-exchange SP-sepharose column and eluted with a linear gradient of 0-100% buffer CE500 (20mM HEPES pH7.6, 500mM NaCl). Peaks fractions were combined and dialysed overnight into CE300 (20mM HEPES pH7.9, 300mM NaCl). Protein was concentrated to ~30mg/ml with a Vivaspin centrifugal 5 kDa cut-off concentrator.

3.2.3 Production of selenomethionyl MBD protein

The recombinant plasmid was transformed into the methionine auxotroph B834(DE3)pLys. A single colony grown on an LB agar was used to inoculate 50ml LB containing 50μg/ml kanamycin. Cells were harvested from 40ml overnight culture by centrifugation at 5000*xg* for 15min at 4˚C and resuspended in 10ml minimal medium [2mM $MgSO_4$, 2X M9 (37.4mM NH_4Cl, 44mM KH_2PO_4, 95.8 mM

Na_2HPO_4 anhydrous), 90mM $FeSO_4.7H_2O$, 0.4 % (w/v) glucose, 1μg/ml vitamin mixture (riboflavin, niacinamide, pyridoxine monohydrochloride, and thiamine at 1μg/ml each), selenomethionine (10 mg/ml) and the other 19 amino acids (4 mg/ml each), 50μg/ml kanamycin] prior to addition into 1L minimal media. Expression and purification of selenomethionyl protein was then performed as for the native protein except that 1mM of DTT was incorporated in all buffers.

3.2.4 Measuring protein concentration

This technique was developed based upon UV adsorption of aromatic amino acids. The protein concentration (mg/ml) is directly proportional to the protein absorbance at 280nm. To determine protein concentration with possible nucleic acid contamination, the spectrophotometer was calibrated to zero absorbance with buffer solution (as blank). Then, A280nm and A260nm of the protein sample were taken and the protein concentration was calculated with equation below (Stoscheck, 1990):

$$Concentration(mg/ml) = (1.55 \times A280_{nm}) - (0.76 \times A260_{nm})$$ **Equation 3-1**

3.2.5 SDS-PAGE

Proteins were separated by SDS-PAGE using a discontinuous buffer system (Laemmli, 1970). SDS-polyacrylamide gels were prepared by using the Mini Protean III apparatus (BioRad). Resolving gel solution [15% (w/v)] containing dH_2O (3.75 ml), 4X lower buffer [1.5M Tris-HCl (pH8.6), 0.4% (w/v) SDS; 3.75ml], bisacrylamide solution [30% (w/v) acrylamide, 0.8% (w/v) bisacrylamide; 7.5ml], 10% (w/v) ammonium persulphate (93.8 μl) and TEMED (BioRad; 10μl) was dispensed into the assembled gel apparatus and the solution surface was levelled with isobutanol (~100μl). The gels were allowed to polymerise for 40 min and the isobutanol layer was removed via absorption with a 3MM Whatman filter paper. The stacking gel solution (5% w/v) containing H_2O (5.84 ml), 4X upper buffer [0.5 M Tris-HCl (pH6.8), 0.4% (w/v) SDS; 2.5ml], bisacrylamide solution [30% (w/v); 1.66ml], ammonium persulphate [10% (w/v); 66.6 μl) and TEMED (BioRad; 7 μl) was then layered onto the resolving gel and a plasmid comb was carefully inserted into the stacking gel solution and the gel was allowed to polymerise for 45 min.

Protein samples (5 µl each) were mixing with 2X loading buffer (62.5 mM Tris-HCl (pH 6.8), 25% (v/v) glycerol, 2% (w/v) SDS, 0.01% (w/v) bromophenol blue, 50mM DTT; 5 µl] and boiled for 5 min. The boiled sample and appropriate protein markers were then electrophoresed with Tris-glycine [0.025 M Tris (pH8.3), 0.192 M glycine, 1 % (w/v) SDS] at a constant voltage (200 mV) until the dye front reached the bottom of the resolving gel (~1 h). The gels were then stained with staining solution [0.25% (w/v) Coomassie blue R-250, 45% (v/v) methanol, 10% (v/v) acetic acid] for 15 min at room temperature with agitation and destained with destaining solution [10 % (v/v) methanol, 10 % (v/v) acetic acid] until the band become clear (~2-3 h).

3.2.6 Western blot

Immunoblotting was performed using the semi-dry Western blot method (Towbin *et al.*, 1979). The polyacrylamide gel containing the fractionated proteins was equilibrated in transfer buffer [25 mM Tris-HCl, 190 mM glycine (pH 8.0), 20% (v/v) methanol)] for 10 min at room temperature. Meanwhile, a piece of nitrocellulose membrane and 6 pieces of Whatman 3MM filter paper, approximately the size of a gel, were soaked in the transfer buffer in a separate container for 5 min at room temperature. The filter papers (3 pieces) were placed on the lower electrode (anode) of the semi-dry blotter (Trans-Blot® SD, BioRad) and the air bubbles were removed by rolling a glass rod on top of the papers. The pre-soaked nitrocellulose membrane was placed on top of the filter papers and the equilibrated gel was transferred carefully onto the membrane. Another 3 pieces of filter paper were placed on the top of the gel and the air bubbles were removed as above. The upper electrode (cathode) and the safety cover were then assembled. Afterwards, a constant current (60mA) was applied for 1-2 h to transfer the proteins onto the nitrocellulose membrane. Lanes containing the molecular weight markers were cut from the membrane and stained with staining solution for 1 min and immediately soaked in destaining solution for 1 h. The membrane was blocked with milk diluent (KPL, USA), which was 10-fold diluted in dH_2O, for 2 h at room temperature. The nitrocellulose membrane was then washed 3 times with TBST [TBS (pH 7.6) supplemented with 0.05% (v/v) Tween 20] for 5 min intervals. The nitrocellulose membrane was incubated with mAb anti-His (Qiagen) (1:5000 diluted in TBST) for 1 h at room temperature with gentle agitation. The membrane was then washed again with TBST, which was followed by incubation

with alkaline phosphatase conjugated goat anti-mouse antibody (1:5000 diluted in TBST) for 1 h at room temperature. The immunoblotted bands were developed with BCIP/NBT in alkaline phosphatase buffer [100mM Tris-HCl (pH 9.5), 100mM NaCl, 5mM $MgCl_2$]. Colour development was terminated by incubating the nitrocellulose membrane with dH_2O and the developed membrane was allowed to dry at room temperature.

3.2.7 Site-directed mutagenesis

A pair of mutagenic primers was designed to contain the desired point mutation. The plasmid bearing the native MBD coding region was used as a template for generation of mutant strands using PCR with *PfuTurbo* DNA polymerase according to the manufacturer's instruction (QuikChange, Stratagene). The parental strands were then digested with Dpn *I* restriction enzyme and the treated plasmid was transformed into XL1-Blue supercompetent cells and plated onto LB agar plate containing kanamycin. After overnight incubation, single colonies were picked and inoculated into 10ml LB supplemented with kanamysin. Plasmids were then isolated with a Miniprep Extraction kit (Qiagen) and the clones were verified by DNA sequencing. Mutant plasmids was then transformed into B834(DE3)pLys cells and selenomethionyl mutant protein was expressed and purified as described above.

3.2.8 End-labelling DNA with gamma ^{32}P-dATP

All oligonucleotides for EMSA were synthesized by Sigma-Genosys and purified by desalting. The labelling reaction containing DNA duplex (50ng), T4 polynucleotide kinase (New England BioLabs) (20 units); γ^{32}P-dATP (Amersham Biosciences) (1 μl); 10X T4 polynucleotide kinase reaction buffer (0.7 M Tris-HCl, 100 mM $MgCl_2$, 50 mM DTT; 3μl) was incubated at 37ºC for 1 h. Free nucleotides in the labeling reaction were removed with nucleotide removal kit (Qiagen) according to the manufacture instruction.

3.2.9 Electrophoretic mobility shift assay (EMSA)

Binding reactions containing 5% (w/v) ficoll-400, 20mM HEPES pH7.9, 100mM NaCl, 25ng/ml polydG-dC.polydG-dC (Amersham Biosciences) competitor DNA and recombinant protein were pre-incubated at room temperature for 10 min. Radio-

labelled probe DNA was then added and the reaction was continued for a further 25 min at room temperature. Sucrose tracking dye [(40% (w/v) sucrose, 0.01% (w/v) bromophenol blue, 0.01% (w/v) xynole xylane blue)] were added and the reaction loaded on to an 8% (w/v) native polyacrylamide gel which had been pre-run at 240V for 30min at 4˚C in 0.5X TBE. The samples were electrophoresed for 3 h under the same condition. Gel was transferred to a 3MM Whatmann paper and dried for 1 h at 80˚C. Dried gels were exposed to Phosphor Imager screen overnight and developed with the Storm machine the following day.

3.2.10 Mass spectrometry

The purified protein was diluted with ammonium phosphate buffer [50mM $(NH_4)_2PO_4$, pH7). Electrospray ionisation mass spectrometry analysis was carried out on a single quadropole mass spectrometer (Micromass, Wyntheshawe, UK) equipped with a Z-spray electrospray ionisation source. Spectra were analysed with a Mass Lynx v3.5 (Micromass, Wyntheshawe, UK).

3.2.11 Gel filtration analysis

The protein and DNA mixture containing the construct 77-167 and 20 bp *BDNF* fragments in a ratio of 1:1.3 was analysed with Superdex 75 HR 10/30 column on an FPLC system (Amersham) at 4˚C

3.3 RESULTS AND DISCUSSION

MeCP2 domains being expressed and purified were constructs 1-205, 78-205 and 77-167 (numbering according to human MeCP2 amino acid sequence) (Figure 1-10 and Figure 3-4). The largest construct comprises of the MBD domain flanked by the N-terminal region (amino acid 1-76) of MeCP2 and a region composed of the AT hook motif (^{185}GRGRGRPK^{192}) at its C-terminus. Construct 78-205 contains the MBD domain and the AT hook motif at its C-terminus and the smallest construct; 77-167; bearing only the MBD domain. To facilitate protein purification with IMAC, an uncleaveable 6xHis tag was genetically fused to the C-terminus of each construct.

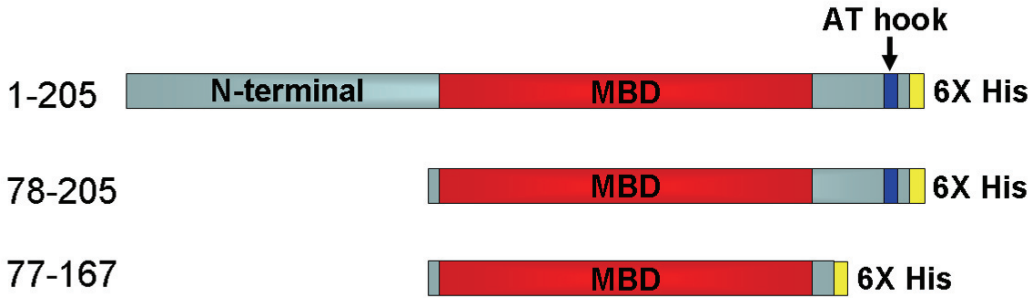


Figure 3-4: MeCP2 constructs used in this study
1-205, the MBD is flanked by the N-terminal region of MeCP2 consisting of 76 amino acids and the AT hook (amino acid sequence ^{185}GRGRGRPK^{192}) at the C-terminal region. 78-205, comprises of MBD domain and the AT hook, and 77-167, contains only the MBD domain. All constructs carrying a 6x His tag at their C-terminus.

3.3.1 Purification of construct 1-205

The largest construct was purified using 3-step chromatography: (i) IMAC with Talon (BDH) or Ni-NTA resin, (ii) size exclusion chromatography with Sephacryl s200HR and (iii) cation exchange chromatography with SP-sepharose. The elution profile using SDS-PAGE analysis after IMAC (Figure 3-5a) revealed that the dominant protein ran approximately 32 kDa on-gel corresponding to the construct 1-205 and co-eluted with several higher/lower molecular weight species. These bands might be the breakdown products of the recombinant protein or the host cell proteins. These impurities however can be reduced by fractionation using sephacryl s200HR. Proteins with molecular weight larger or smaller than 32 kDa are separated from the dominant band (Figure 3-5b). Subsequently, most of the remaining impurities can be removed by SP-sepharose chromatography (Figure 3-5c). Although the elution profile of SP-sepharose showed traces of impurities, the protein is approximately 95% pure.

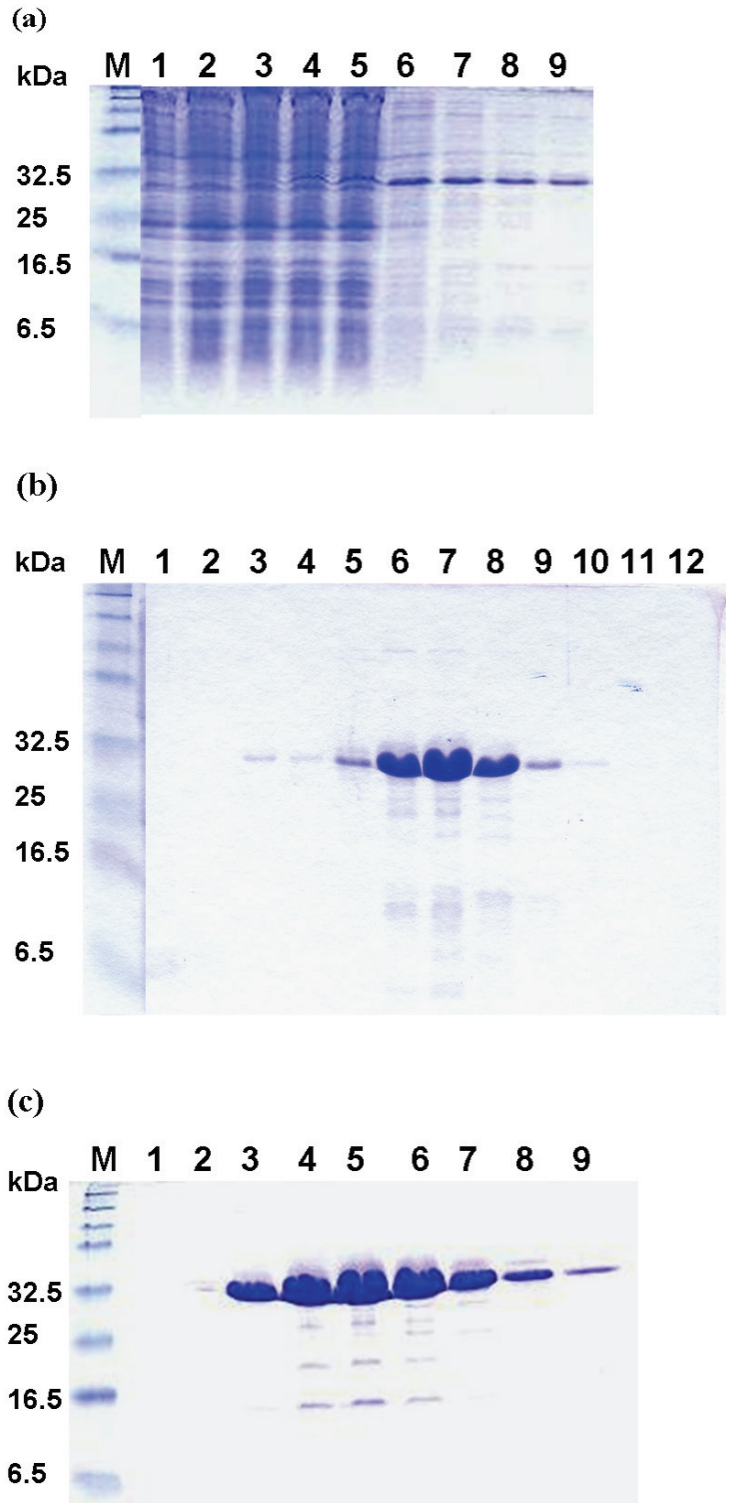


Figure 3-5 SDS-PAGE analysis of construct 1-205
(a) Fractionation of Ni-NTA column showing that the flow through (lanes 1 – 5) contains only traces of protein corresponding to 32 kDa while eluted fractions (lanes 6 – 9) contain mainly this species. (b) Fractionation of Sephacryl s200HR showing that the contaminants larger or smaller than 32kDa were separated from the dominant band and (c) Elution profile after SP-sepharose indicating that the construct 1-205 is approximately 95% pure. Lane M, protein ladder.

3.3.2 Construct 78-205

Similarly, construct 78-205 was also purified with 3-step column chromatography as the largest construct 1-205. Eluted fractions analysed using SDS-PAGE contain a dominant band, approximately 16 kDa, which coexists with other contaminants (Figure 3-6a). After size exclusion chromatography with Sephacryl s200HR, most of these impurities were separated into early and later fractions of the predominant band (~ 16 kDa) (Figure 3-6b). The third column however removed almost all contaminants at least to an undetectable level (Figure 3-6c). The same strategy was used to purify the mutant proteins (T158M/A/S) for mutational analysis.

3.3.3 Construct 77-167

To purify the shortest construct used in this study, 2-step column chromatography was sufficient to give highly pure protein for crystallisation and biophysical analysis. The SDS-PAGE profile from fractions eluted from Ni-NTA/Talon IMAC shows a reasonably pure predominant band in between 6.5 and 16.5 kDa together with other impurities (Figure 3-7a). The second column using Sephacryl s200HR exclusively removed all contaminants (Figure 3-7b). For mutational analysis, the mutants created using site directed mutagenesis approach were T158M, T158S, T158A and Y123F of 77-167. All mutants were purified using the same method.

In order to solve the phase problem, seleno-Met derivatives (wildtype and A140M of 77-167) were overexpressed and purified. All buffers used were supplemented with 1mM DTT in order to prevent oxidation of selenium attached covalently to the protein. Unfortunately, not all IMAC resins are compatible with reducing agent such as DTT or β-mercaptoethanol. In the presence of 1mM DTT, Ni^{2+} attached to NTA resin remained in the oxidised form whereas Co^{2+} in Talon is reduced to thiol complexes which destablises the resin. Therefore, Ni-NTA IMAC was the method of choice for purification of seleno-Met derivatives. An additional 'wash' step is required after sample loading to remove the weak binders. The rest of the purification was exactly the same as the wild type protein except in the presence of 1mM DTT. The purified proteins were analysed with mass spectrometry and Western blotting. Gel shift assays were carried out to characterise their DNA binding capability and to provide a rough K_d determination.

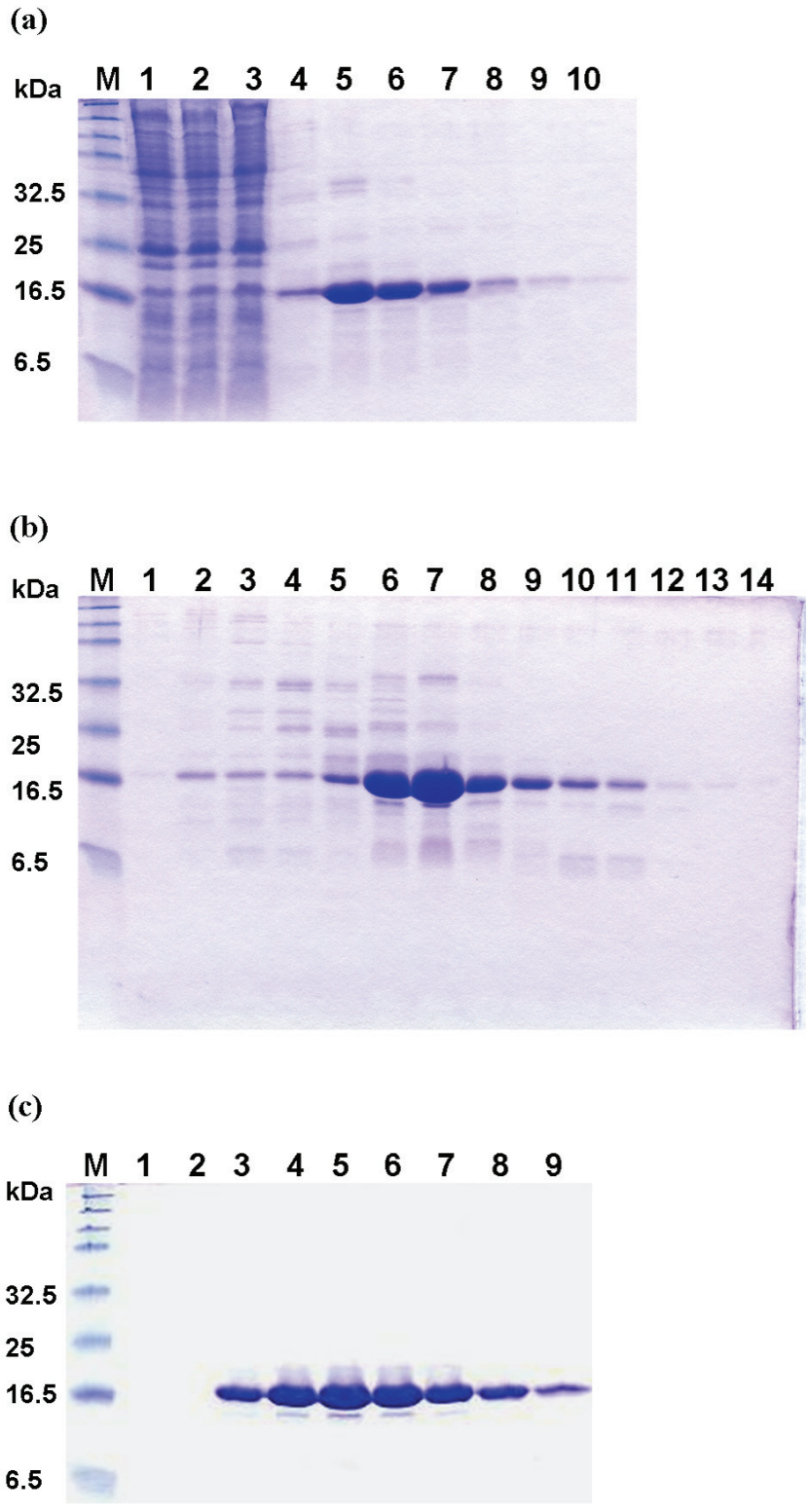


Figure 3-6: SDS-PAGE analysis of construct 78-205
(a) Fractionation of Ni-NTA column showing that the flow through (lanes 1 – 3) contains only traces of protein corresponding to 16.5 kDa while eluted fractions (lanes 4 – 10) contains mainly this species. (b) Elution profile after Sephacryl s200HR showing that the contaminants larger or smaller than 16.5 kDa were separated from the dominant band and (c) Elution profile after SP-sepharose indicating that the construct 78-205 is above 95% pure. Lane M, protein ladder.

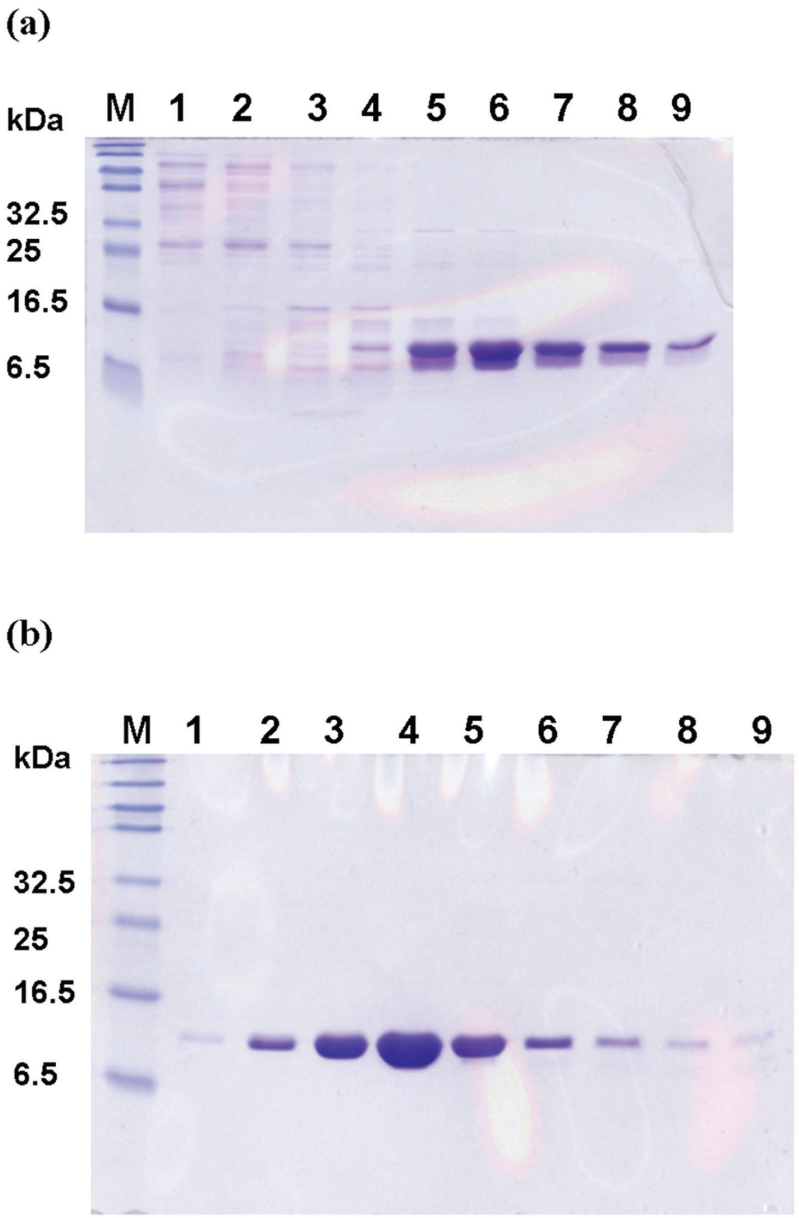


Figure 3-7: SDS-PAGE analysis of construct 77-167
(a) Elution profile after Talon column showing that all fractions contains a protein with molecular weight in between 6.5 to 16.5 kDa and (b) Elution profile of Sephacryl s200HR showing that the construct 77-167 is nearly 100% pure. Lane M, protein ladder.

3.3.4 Preliminary characterisation

3.3.4.1 Mass spectrometry

SDS-PAGE was initially used to estimate the molecular weight of MeCP2 contructs. The molecular weight estimated from the SDS-PAGE, however, could not represent the actual molecular mass of the proteins. These on-gel estimated values (Table 3-1)

are distinct from theoretical molecular weight. The primary sequence of constructs 77-167, 78-205 and 1-205 contain 18.5% (18 residues), 22.4% (30 residues) and 21% (44 residues) positively charged amino acids (Arg and Lys), respectively. The net positive charged of constructs 78-205 and 1-205 is enriched by Arg-rich C-terminus particularly the AT hook of MeCP2. Positively charged of construct 1-205 is further enhanced by 13 Lys residues at the N-terminal region of MeCP2. The high content of basic amino acid leads to a predicted isoelectric point approximately 10. The high concentration of positively charged amino acids may account for the anomalous mobility of MeCP2 constructs in SDS-PAGE. Therefore, mass spectrometry was carried out to determine the accurate mass of these constructs. In principle, proteins in solution are transformed into ions (as an intact ionised molecule) *in vacuo* using an ion source and these ions are separated by their *mass-to-charge* ratio (*m/z*) in an electric or magnetic fields. These ions can then be analysed quantitatively and qualitatively by their *m/z* values and abundance.

Table 3-1 Mass spectrometry

Construct	pI	Molecular weight / Da		
		On-gel	Theoretical	Mass spectrometry
77-167	9.99	Between 6500 -16500	11103.4	11108.6
78-205	10.63	~16500	14997.8	14978.5
1-205	9.88	~32500	23243.1	23152.0

Theoretical molecular weight and pI values of these constructs were calculated using the Protparam tool at http://www.expasy.ch/tools/protparam.html. A pI value is a pH where the protein has a net charge equal to zero.

3.3.4.2 Western blot

Western blot was carried out to analyse the purity of each purified and concentrated MeCP2 construct. Proteins were separated on an SDS-PAGE gel, transferred onto nitrocellulose membrane and immunoblotted with anti-His monoclonal antibody. Predominant bands corresponding to 12 kDa, 16 kDa and 32 kDa (Figure 3-8; lane 1, 2 and 3), respectively, were observed in lanes containing 77-167 (lane 1), 78-205 (lane 2) and 1-205 (lane 3) construct preparations. These bands were successfully immunoblotted with mAb anti-His and therefore might correspond to 6xHis-tagged recombinant MeCP2 domains. However, a small amount of other bands below/above the predominant ones; which can be detected with mAb anti-His; have also been

observed. They were thought to be the mildly truncated MeCP2 domains formed in the process of purification or bacterial host proteins that can be detected with mAb anti-His.

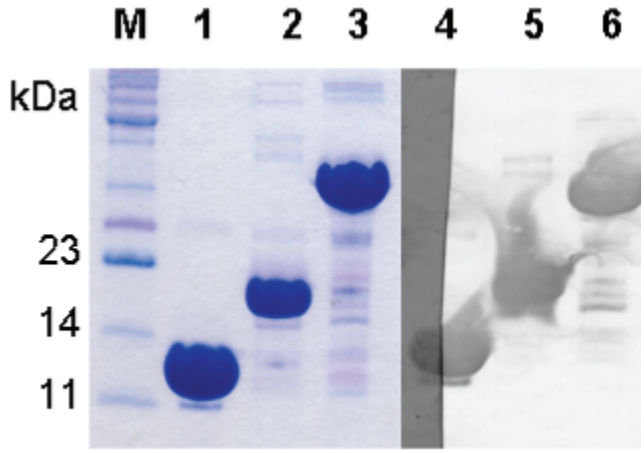


Figure 3-8: Western blot of concentrated MeCP2 proteins.
The separated proteins were stained with coommasie blue (lanes M – 3), and transferred to a nitrocellulose membrane and blotted against mAb anti-His (lanes 4 -6).

3.3.4.3 Electrophoretic mobility shift assay (EMSA)

The first step toward the co-crystallisation of DNA-protein complexes is to select a stable DNA-protein complex formed under a physiological condition. The DNA-protein complexes can then be analysed using various methods such as EMSA, filter binding assays and fluorescence binding assay to characterise their binding capability. Oligonucleotides (probe S1 and S2) used in this study were chosen from methyl-SELEX experiments reported in Klose *et al.* (2005). Both DNA sequences have been truncated to 19 bp to contain a central methyl-CpG pair and an AT run. The AT run of these sequences is located 3 bases apart from the methyl-CpG. In addition, a 19 bp DNA fragment corresponding to the *BDNF* promoter was also chosen for a preliminary DNA-protein binding assay because this DNA fragment also contains a methyl-CpG and an AT run. Most importantly, the *BDNF* promoter is an endogenous MeCP2 binding target (Chen *et al.*, 2003; Martinowich *et al.*, 2003). The methyl-CpG is separated from the AT run by only a single base pair. The truncation of these oligonucleotides to a relatively short length is coherent with the purpose to design a rational length for DNA-protein co-crystallisation. The MeCP2 constructs used in this assay include all three constructs previously purified.

The EMSA was carried out with two different protein concentrations; at 500 and 1000nM; and DNA probes labelled with γ-P^{32} dATP. Figure 3-9 shows that all three oligonucleotides bind tightly to the shortest construct 77-167 over 78-205 and 1-205. The smearing effect in the lanes containing 78-205 and 1-205 indicates that the DNA-protein complexes were not stable and dissociated while migrating in the polyacrylamide gel. Among the DNA duplexes, probe S1 demonstrates the strongest protein binding capability, followed by S2 and *BDNF* fragments. A possible explanation for the weak binding of 1-205 is that the flexible long N-terminal tail interrupts the methyl-CpG binding region of the MBD domain and the DNA is hindered from binding efficiently onto its target site. In the absent of the N-terminal region, only a small amount of 78-205 was involved in methylated DNA binding. The length of the DNA might not be long enough to accommodate the C-terminal region (including the AT hook). One possibility is that in the absence of its cognate binding region the AT hook might be relatively unstructured and this destabilises the DNA-protein association. The DNA-dependent stabilisation of the AT hook indicates a longer DNA is required to form a stable DNA-protein complex. The presence of an AT run on the DNA might influence the groove width which is possibly involved in protein-DNA interaction. Together with the DNA primary sequence, the major difference between the 3 methylated DNA is the gap between the methyl-CpG and the AT run. It is unknown how this disparity influences the DNA binding.

Although the protein binding efficiency of oligonucleotides S1 and S2 is higher than the *BDNF* fragment, the later was chosen to carry out the initial co-crystallisation screening. The main reason is that *BDNF* fragment is part of the *BDNF* promoter region, which has been identified as an endogenous target gene of MeCP2 (Chen *et al.*, 2003; Martinowich *et al.*, 2003). Artificially selected S1 and S2 using the methyl-SELEX experiment might not correspond to a biologically active DNA sequence. However, the methyl-CpG and the newly identified AT run characteristics explained the DNA binding specificity of MeCP2 and highlights the importance of the AT run in MeCP2 cognate binding targets. Out of the three MeCP2 constructs, 77-167 forms the most stable complex with the *BDNF* fragment. Therefore, this construct was chosen for DNA-protein crystallisation trials.

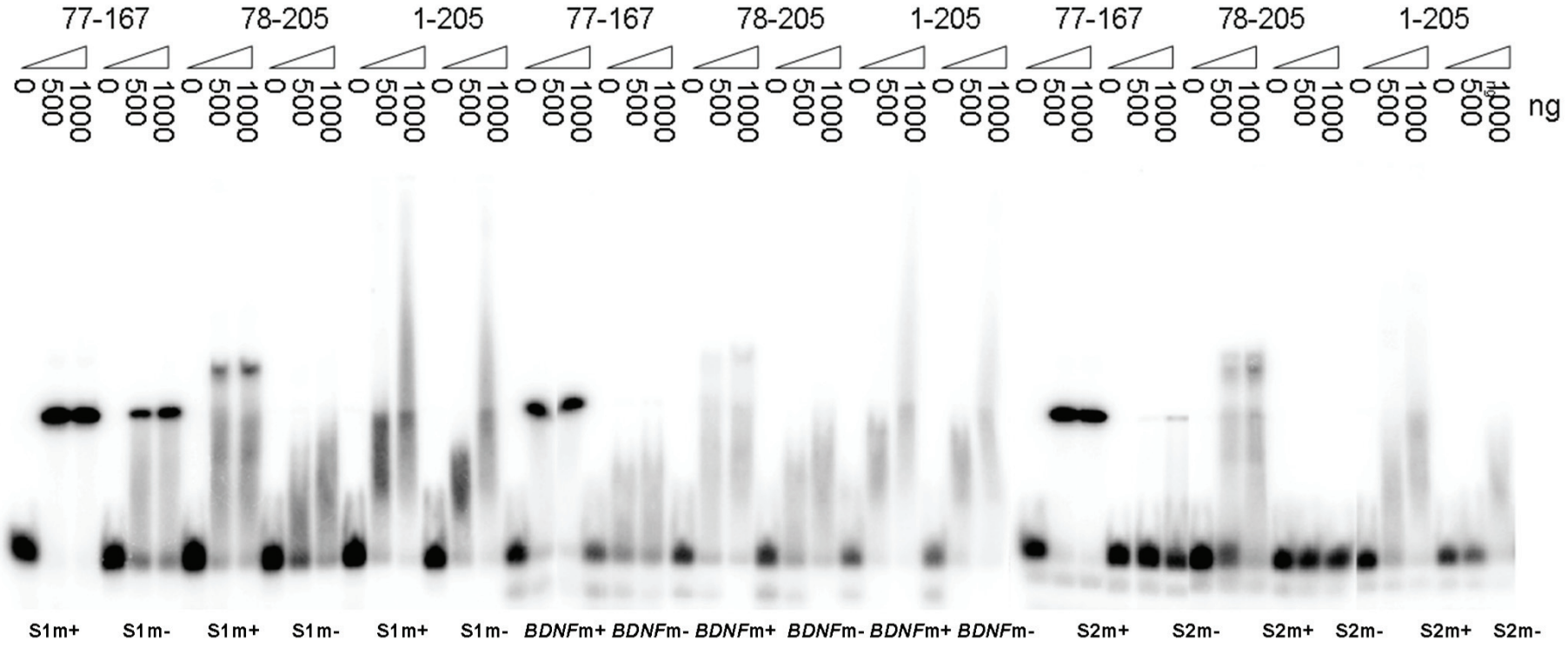


Figure 3-9: Preliminary characterisation with EMSA.
Methyl-Selex determined oligonucleotides (probes S1 and S2) and *BDNF* sequence (*BDNF*), both methylated (m+) and unmethylated (m-), were end-labelled with γ-P^{32} –dATP assayed with three different constructs used in this study. All methylated probes bound tightly to construct 77-167 (bearing only MBD domain) but weakly to constructs 78-205 and 1-205. Primary sequence of S1, S2 and *BDNF* probes are: GGGGGC(m5C)GGAGTTAAGGAC, CGACCC(m5C)GGAGATAACG and CTGGAA(m5C)GGAATTCTTCTA, respectively.

3.3.5 Quantitative EMSA

To measure the binding affinity, a quantitative EMSA analysis using 77-167 and probes S1, S2 and *BDNF* fragments was carried out (Figure 3-10). Larger constructs 78-205 and 1-205 were excluded from this assay due to their weak binding affinity to the DNA fragments. The K_d values for the binding data were calculated with non-linear hyperbolic curve fitting method as described by Valinluck and coworkers (Valinluck *et al.*, 2005; Valinluck *et al.*, 2004). Assuming that the MBD of MeCP2 binds to the methylated DNA substrate as a monomer (Nan *et al.*, 1993), the equilibration of MBD and methylated DNA can be written as:

$$[\mathrm{E}]+[\mathrm{O}] \rightleftharpoons [\mathrm{EO}]$$

Where [*E*] is unbound MBD concentration, [*O*] is unbound duplex concentration, and [*EO*] is MBD-DNA complex concentration. Using this single binding scheme, the dissociation constant (K_d) can be defined as:

$$\mathrm{K_d} = \frac{[\mathrm{E}][\mathrm{O}]}{[\mathrm{EO}]} \qquad \textbf{Equation 3-2}$$

The equation can be rearranged to give:

$$[\mathrm{O}] = \frac{\mathrm{K_d}[\mathrm{EO}]}{[\mathrm{E}]} \qquad \textbf{Equation 3-3}$$

The fraction bound was plotted as a function of protein concentration. The Fraction duplex bound can be calculated as:

$$\text{Fraction duplex bound} = \frac{[EO]}{[EO]+[O]} \qquad \textbf{Equation 3-4}$$

When unbound protein is in excess ([*E*] > [*O*], then $[E]_{total} - [EO] = [E] \approx [E]_{total}$; where $[E]_{total}$ is the total MBD concentration including the bound and unbound protein. Substitution of Equation 3-3 into Equation 3-4 yields:

$$\text{Fraction duplex bound} = \frac{[\mathrm{E}]_{total}}{[\mathrm{E}]_{total}+\mathrm{K_d}} \qquad \textbf{Equation 3-5}$$

The dataset was fitted into the equation by non-linear regression using Grafit software. Among the three methylated DNAs, methylated S1 shows the highest affinity (70.9 nM) for construct 77-167, which is approximately 15-fold higher than that of the *BDNF* fragments (Table 3-2). The poorly fitted curve of the *BDNF* fragment roughly determined the K_d at 1μM under this assay condition. Among these DNA, duplex S1 binds strongest to MBD, which followed by S2 then *BDNF*.

Table 3-2 K_d of construct 77-167 binding to 19 bp DNA diplex

Duplex	K_d (nM)
BDNF	1039.5 ± 352.0
S1	70.9 ± 25.6
S2	88.5 ± 33.3

These K_d values determined in this experiment can only be used as a guide to distinguish the relative binding affinity between the DNA fragments. These values are inappropriate to represent the actual K_d. Several criteria were not fulfilled in this K_d determination. These include: (i) the temperature was not kept constant throughout the experiment. The DNA-protein mixtures were incubated at room temperature to promote complex formation but electrophoresis was performed at 4 ˚C; (ii) re-association of DNA-protein cannot occur once the mixtures migrated into the gel because the unbound DNA, unbound MBD and the complexes travelled at different rates. Thus, association and dissociation equilibrium cannot be achieved in EMSA and; (iii) the binding curve for the *BDNF* fragment has not reached a saturation state at the highest protein concentration and; (iv) only one measurement was taken. The actual K_d can be determined in a more precise way using isothermal titration calorimetric (ITC) and surface plasmon resonance (SPR).

Several binding affinities of symmetrically methylated DNA and MeCP2 primarily using EMSA have been reported (Free *et al.*, 2001; Nan *et al.*, 1997; Valinluck *et al.*, 2005; Valinluck *et al.*, 2004). Using the MBD fragment of MeCP2 and 27 bp methylated DNA, a K_d value of 15 nM has been determined whereas the binding affinity reduced to 1 μM with methylation free DNA (Valinluck *et al.*, 2005; Valinluck *et al.*, 2004). These reports focused on the recognition of symmetrical methyl-CpG by MeCP2 MBD domain under various experimental conditions. Before the finding of the AT run requirement, the importance of flanking DNA sequences was down-weighted. Based on this recent discovery (Klose *et al.*, 2005) and newly identified MeCP2 target genes such as *BDNF*, and *Dlx6* (both contain A/T bases close to methyl-CpG), new assays must be developed to quantitatively measured the binding affinity of MeCP2 on the methylated DNA bearing A/T sequences.

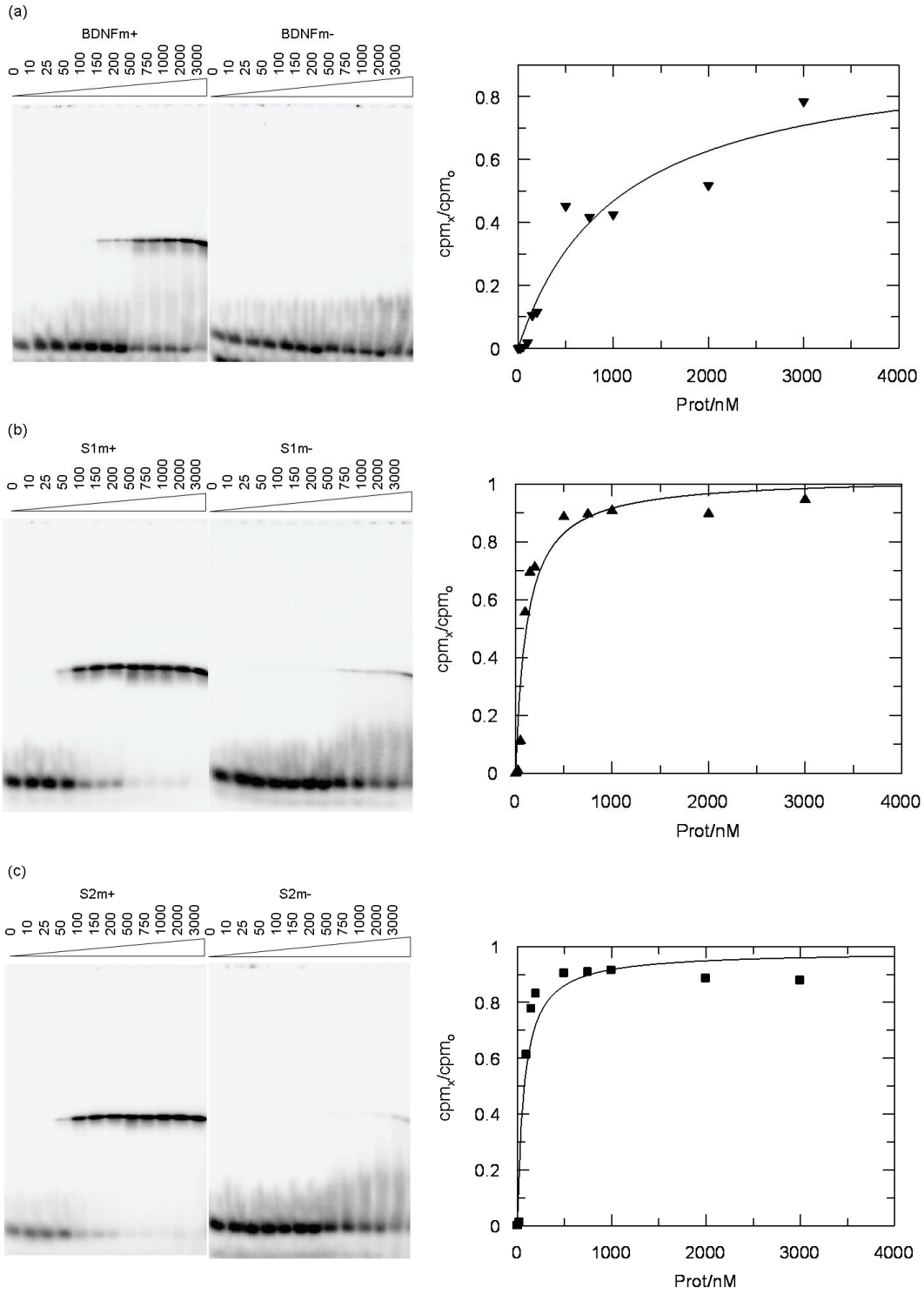


Figure 3-10 K_d determination

Left panel; binding of (a) *BDNF*, (b) S1 and (c) S2 duplexes to varying concentrations of 77-167 from 0 to 3 μM.

Right panel; non-linear regression of the plots of percentage binding of (a) *BDNF*, (b) S1 and (c) S2 duplexes versus concentration of construct 77-167. The x-axis represents MBD concentration (nM) and the y-axis represents fractions of probes binding. m+ and m- indicate methylated and unmethylated probes, respectively.

3.3.6 Gel exclusion analysis

In order to prepare a DNA-protein complex for co-crystallisation, construct 77-167 and 20 bp *BDNF* fragments were mixed in a ratio 1:1.3 in buffer CE100 (20mM HEPES pH7.6, 100mM NaCl). The mixture was incubated for 30 min at room temperature to promote DNA-protein complex formation. To ensure the complex formation, a diluted sample was analysed using a Superdex 75 HR10/30 column. Figure 3-11 shows the elution profile of the gel exclusion analysis. A total of two peaks were observed with the first peak height approximately 7.7-fold higher than the second peak. Because the DNA in the mixture was 30% in excess, the result suggested that all proteins molecules were complexed with the added DNA while the excess DNA was eluted as a second peak.

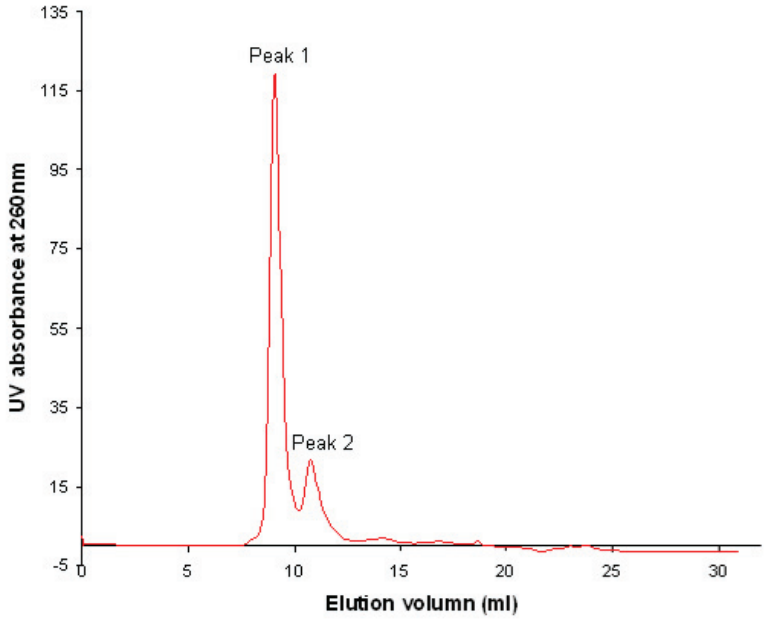


Figure 3-11 Gel exclusion analysis
The DNA-protein complex was prepared by mixing the MBD and *BDNF* fragment in 1:1.3 ratio. Peak 1 was eluted immediately after the void volume of Superdex 75 HR 10/30 column (7.7 ml) followed by peak 2. According to the calculated ratio, peak 1 and peak 2 correspond to the DNA-protein complex unbound duplexes as the ratio of height of these peaks is approximately 1:0.3.

3.4 SUMMARY

MeCP2 domains (77-167, 78-205 and 1-205) were successfully expressed and purified. Construct 77-167 can be purified using Ni-NTA column and gel filtration (Hi-prep 16/10 Sephacryl 200HR) to achieve homogeneity. In contrast, constructs 78-205 and 1-205 require an additional step (SP-sepharose) to achieve purify greater than 90%. Preliminary characterisation using a gel shift assay demonstrated that the smallest construct binds strongly to methylated probes S1 and S2 (Methyl-SELEX probes) but relatively weaker to methylated *BDNF* sequence. By considering the chances to crystallise a protein-DNA complex, the smallest construct 77-167 (MBD domain) and a short fragment of *BDNF* promoter sequence, which has been identified as the MeCP2 target gene, were chosen for co-crystallisation screening. The K_d values obtained in this study are relative values of DNA binding affinity between probes and cannot represent the actual dissociation constant because several requirements are not fulfilled. Gel filtration analysis using Superdex 75 HR 10/30 column shows that the ratio of DNA-protein complex to free DNA is approximately 1:0.3, which clearly represents the actual components in the protein-DNA mixture. As a result, DNA-protein co-crystallisation trials were carried out using various fragments of *BDNF* sequence and construct 77-167.

CHAPTER 4. CRYSTALLISATION AND STRUCTURAL DETERMINATION

4.1 INTRODUCTION

To crystallise a DNA-protein complex, there are more considerations than for the case of crystallising the protein alone. In addition to the general considerations for protein sample preparation such as buffer choice, concentration and purity, the experimenter must consider the length and terminal nucleotides of the DNA and how the complex is formed. Double stranded DNA has a strong preference to stack end-to-end in the crystal packing, therefore, the precise length of the DNA fragment and the nature of the stacking interactions between fragments are important determinants of the unit cell and crystalline order. The length of DNA used in the DNA-protein complex co-crystallisation is usually beyond the minimal DNA sequence recognised by the protein. Co-crystallisation trials can begin from the minimum DNA sequence required for protein binding and gradually increase the number of sequences. Lengths of the DNA particularly successful in crystallisation trials usually correspond to the multiples of integral or half integral turns of the DNA.

Another crucial consideration is the composition of DNA ends. Single or double complementary overhanging bases at either the 3' or 5' end are commonly included to promote end-to-end stacking of oligonucleotides in the crystal although blunt-ended oligonucleotides have been reported in a number of successful cases. Among the different types of overhangs that have been successfully used, the most frequent one is a single 5' overhanging A on one strand and a 5' T overhanging on the other. The DNA used in crystallisation can be synthesised on a 1 μmole scale and dissolved to give 1 mM duplexes. Single stranded DNA should be purified by reversed-phase HPLC to purity greater than 90%.

The DNA-protein complex is prepared by adding the concentrated protein solution directly but slowly to the DNA solution. It is usual to have a slight excess of DNA (typically 10 to 20 %) rather than using a 1:1 ratio of DNA to protein. Preliminary characterisations such as band shift assay should be carried out to determine the actual stoichiometry of the mixture if this has not been done in previous biophysical assays.

As mentioned in Chapter 1, the NMR established MBD domain of MeCP2 displays a wedge shape structure composed of four anti-parallel β-strands and one α-helix (Wakefield *et al.*, 1999). Some residues located at loop L1 (Gly114 and Ala117), β3 (Ala131), loop L2 (Arg133), α1 (Lys135, Val136), and the C-terminal region (Phe157, Thr160 and Arg162) showed amide proton chemical shift changes upon addition of methylated DNA. However, none of the amino acids of β2 and β3 undergo significant chemical shift changes. Recently, Klose et al (2005) discovered that an AT run adjacent to the methyl-CpG is required for high affinity binding of MeCP2. Interestingly, endogenous target genes such as *BDNF* and *Dlx6* also contain AT run(s) close to the mCpG (Klose *et al.*, 2005). In contrast to MeCP2, MBD1 does not require an AT run to promote tight DNA binding (Klose *et al.*, 2005). Therefore, the binding specificity of MeCP2 is distinct from MBD1. Moreover, mutations of MeCP2 are responsible for RTT with higher than 50% occurrence clustered within the MBD domain (Kriaucionis and Bird, 2003). Therefore, a high resolution structure of MeCP2 in complex with methylated DNA is required to explain the molecular details particularly the methyl-CpG recognition in the major groove of the DNA.

The first part of this chapter describes the co-crystallisation of the MeCP2 MBD domain in complex with a 20 bp DNA fragment of promoter III (nucleotides -108 to -90) of the *BDNF* gene which contains a central methyl-CpG pair and an AT run. The second part of this chapter presents the strategies used in solving the X-ray structure of MeCP2 MBD complexed with methylated DNA. Initially, cocrystals of wild type MeCP2 MBD-DNA complex were obtained but an X-ray structure was not solved due to insufficient phase information. Later, an MBD mutant containing A140M was created by site-directed mutagenesis. seleno-methionyl protein was prepared and crystals containing an additional SeMet (A140SeMet) were successfully grown. From this crystal, the additional selenium signals enable several MAD datasets to be collected at station BM14, ESRF, Grenoble, France. However, the crystal structure of the MeCP2 MBD domain in complex with methylated DNA was eventually determined with the single-wavelength anomalous dispersion (SAD) method using the peak dataset of the selenium derivative.

4.2 MATERIAL AND METHODS

4.2.1 Nucleic acid preparation

Various DNA constructs were designed based on the DNA sequence of the mouse *BDNF* promoter (Chen *et al.*, 2003; Martinowich *et al.*, 2003) which is situated upstream at -148 bp from the transcription start site of an endogenous MeCP2 target region. All oligonucleotides for cocrystallisation were synthesized and HPLC purified by Oligos *Etc*. (Wilsonville, Oregon, USA). The lyophilized oligonucleotides were resuspended in TEN buffer (10mM Tris-HCl pH8.0, 0.5mM EDTA, 100mM NaCl) to a concentration of 2mM. Complementary strands were mixed in equimolar amounts and annealed by heating to 368K and then cooled to room temperature over 3 hours.

4.2.2 DNA-protein complex preparation

The complex of DNA-protein was prepared by mixing 1:1.3 ratio of protein to DNA in CE100 (20mM HEPES pH7.9, 100mM NaCl) and the mixture was incubated at room temperature for 30 min to promote complex formation. The final concentration of the DNA and protein were 260 μM and 200 μM, respectively.

4.2.3 Hanging drop vapour diffusion

An equal volume of the DNA-protein complex and crystallisation solution were mixed and equilibrated against the crystallisation solution using a hanging drop vapour diffusion method (see details in Results and Discussion). Rectangular crystals with maximum dimension of 0.5mm grew within 3 days. The crystals were flash frozen in liquid nitrogen. Most crystals diffracted to ~3Å maximum resolution, but soaking of manganese chloride (10mM) for 15 min successfully improved the resolution to 2.5Å with a significant change in unit cell dimensions.

4.2.4 Microseeding/streak seeding

A cluster of crystals was crushed with a rounded glass rod after transferring to a crystallisation reagent with higher precipitant concentration. If necessary, the seed stock was serially diluted into the same mother liquor. A cat whisker was then drawn across the seed stock and streaked onto a pre-equilibrated drop containing an equal volume of sample and crystallisation reagent, in which, the precipitant concentration

should not promote spontaneous nucleation. Observation was then carried out the following day. Alternatively, crystals in the drop were gently touched with a cat whisker and streak seeded onto overnight equilibrated drops. Newly formed crystals were usually visible within 24 h.

4.2.5 Manganese soaking

A crystal of A140SeMet in complex with methylated DNA was transferred from the mother liquor to a temperature pre-equilibrated solution containing 35% (w/v) PEG 2000, 200mM ammonium acetate, 10mM manganese chloride and 50mM sodium cacodylate pH6.5 and incubated for 15 min before flash frozen in liquid nitrogen.

4.2.6 Data collection and processing

Data collections were carried out at station BM14, ESRF, Grenoble, France or station MAD10.1 SRS, Daresbury, UK. All data presented in Table 4-3 were collected at 100K. MAD data were collected from 3 wavelengths corresponding to selenium peak (PK), inflection point (IP), high remote (HR). SAD data for iodine were collected at Station MAD10.1 SRS Daresbury and Station BM14 ESRF Grenoble, France to maximise the anomalous signal of iodine atom. All data were indexed and integrated using the MOSFLM and scaled with SCALA (Potterton *et al.*, 2002). Special considerations of data processing are detailed in the results section.

4.2.7 Molecular phasing, model building and refinement

Strategies used in experimental phasing and molecular replacement will be covered in the results section.

4.3 RESULTS AND DISCUSSION

4.3.1 Co-crystallisation

4.3.1.1 Wild type MeCP2 MBD complexed with methylated DNA

Based upon the preliminary characterisation with the band shift assay using different DNA isolated from methyl-SELEX experiments (Klose *et al.*, 2005) and various MeCP2 domains, the construct bearing only the MBD domain (construct 77-167) was chosen for co-crystallisation with the *BDNF* promoter sequence. Table 4-1 shows the

sequences of DNA duplexes used in the initial screening of crystallisation conditions. Oligonucleotides ranging from 15 bp to 21 bp either single overhanging bases or blunt-ended have been synthesised and reversed-phase HPLC purified to purity greater than 90 %. These oligonucelotides were initially dissolved in TEN buffer to a concentration of 2mM, then mixed equi-molar to yield 1mM DNA duplexes. The protein was added slowly to the DNA solution to a final ratio of 1:1.3. Because the DNA solution contained low salt concentration (100mM), the protein immediately precipitated upon addition to the DNA solution. All protein precipitate however re-dissolved upon complex formation with the DNA. In order to maximise the DNA-protein complex formation, the mixture was then incubated for 30 min at room temperature. Before setting up the crystallisation plate, the solution was centrifuged at 13,000*g* at 4˚ for 15 min to remove any insoluble materials. This step reduces the spontaneous nucleation rate as the insoluble materials like dirt can potentially promote undesirable nucleation which leads to high number of microcrystals. The crystallisation trials were then set up using the hanging drop vapour diffusion method and equilibrated to 17 or 4˚C.

The initial screening using the 15, 16 and 18 bp duplexes (both single-base overhangs and blunt ended) complexed with construct 77-167 did not produce any DNA-protein co-crystals with the Natrix sparse matrix screens (Hampton Research). Increasing the length of the DNA to 20 bp and 21 bp (Table 4-1), however, successfully yielded crystals from the DNA-protein complexes containing the DNA duplexes with single-base overhangs at either 4 or 17˚C from Natrix screens. The positive hits were obtained from conditions 25, 27 and 48 (Table 4-2). The MBD complexed with blunt ended DNA duplexes also produced microcrystals, nevertheless, optimisations with various parameters such as complex concentrations, temperature (17 °C or 4 °C), additives, and metal ions have not succeeded to improve crystal size and quality.

Table 4-1 Summary of oligonucleotides (*BDNF*) used for co-crystallisation trials

No.	Size	Ends	DNA sequence	Protein	Crystal form	Resolution limit (Å)	Crystal name	Dataset
1	21	overhangs	TCTGGAA^{m}**C**GGAATTCTTCTA GACCTTG^{m}**C**CTTAAGAAGATA	WT	Cluster	~20 (~9Å with microseeding)	-	-
2	21	Blunt	CCTGGAA^{m}**C**GGAATTCTTCTAA GGACCTTG^{m}**C**CTTAAGAAGATT	WT	Small crystals	-	-	-
3	20	Blunt	CTGGAA^{m}**C**GGAATTCTTCTAA GACCTTG^{m}**C**CTTAAGAAGATT	WT	Small crystals	-	-	-
4	20	overhangs	TCTGGAA^{m}**C**GGAATTCTTCTA GACCTTG^{m}**C**CTTAAGAAGATA	WT	Crystals	2.7	*Native*	*Native*
				SeMet94-MBD	Small crystals	~9	-	-
				A140SeMet	Crystals	3.0 2.5	*A140SeMet* *A140SeMet-Mn*	*MAD1* *MAD2*
5	20	overhangs	TC^{I}**U**GGAA^{m}**C**GGAATTCT^{I}**U**CTA G ACCTTG^{m}**C**CTTAAGA AGATA	WT	Crystals	3.2	*Iodo*	*Iodo*
				SeMet94-MBD	Crystals	2.85	*IodoSe*	*IodoSe1* *IodoSe2*
6	20	overhangs	TC^{I}**U**GGAA^{m}**C**GGAATTCTTCTA G ACCTTG^{m}**C**CTTAAGAAGATA	WT	Crystals	2.8	*D22*	*Iodo3*
7	20	overhangs	TCTGGAA^{m}**C**GGAATTCT^{I}**U**CTA GACCTTG^{m}**C**CTTAAGA AGATA	WT	Crystals	2.7	*D21*	*Iodo17*
8	20	overhangs	TCTGGAA^{m}**C**GGAATTCT^{Br}**U**CTA GACCTTG^{m}**C**CTTAAGA AGATA	WT	Small crystal	-	-	-
9	18	overhangs	TGACCC^{m}**C**GGAGATAACGA CTGGGG^{m}**C**CTCTATTGCTA	WT	No crystal	-	-	-
10	18	Blunt	CGACCC^{m}**C**GGAGATAACGA CCTGGGG^{m}**C**CTCTATTGCTA	WT	No crystal	-	-	-
11	16	overhangs	TGGAA^{m}**C**GGAATTCTTC CCTTG^{m}**C**CTTAAGAAGA	WT	No crystal	-	-	-
12	16	Blunt	TGGAA^{m}**C**GGAATTCTTC ACCTTG^{m}**C**CTTAAGAAG	WT	No crystal	-	-	-
13	15	overhangs	TGGAA^{m}**C**GGAATTCTT CCTTG^{m}**C**CTTAAGAAA	WT	No crystal	-	-	-
14	15	Blunt	GGAA^{m}**C**GGAATTCTTC CCTTG^{m}**C**CTTAAGAAG	WT	No crystal	-	-	-

^{m}**C**, 5'methyl-cytosine; ^{I}**U**, 5'iodo-uracil; ^{Br}**U**, 5'bromo-uracil. The types of DNA ends are indicated as overhangs (single-base overhangs) and blunt (blunt ended). Dashes indicate no crystal observed or no diffraction. Proteins are indicated as WT (wild type), and SeMet (seleno-Met MeCP2 MBD). The diffraction limit and dataset collected for each crystal are indicated. Statistics for data processing (with resolution better than 3Å) are presented in Table 4-3.

Table 4-2 Positive hits from Natrix screens (Hampton Research)

Condition	Precipitant	Buffer	pH	Ions
25	30% (w/v) PEG4000	50mM sodium cacodylate	6.5	80mM magnesium acetate
27	30% (w/v) PEG8000	50mM sodium cacodylate	6.5	200mM ammonium acetate, 10mM magnesium acetate
48	30% (w/v) PEG4000	50mM Tris-HCl	8.5	200mM ammonium chloride, 10mM calcium chloride

These conditions yielded microcrytals from 20 and 21 bp DNA complexed with protein construct 77-167.

The best optimised condition for single-base overhangs 21 bp DNA complexed with the wild type protein was 30-33% (w/v) PEG4000, 200mM NH_4Cl, 10mM $MgCl_2$ and 50mM Na cacodylate pH5.5 – 7.0 (oligos number 1, Table 4-1). However, spontaneous nucleation with this condition produced low quality crystals for X-ray diffraction (maximum resolution ~20Å) (Figure 4-1a). Microseeding using these crystals as seeds improved the crystal appearance and resolution to 9Å (Figure 4-1b). Further rounds of seeding however did not improve the resolution limit.

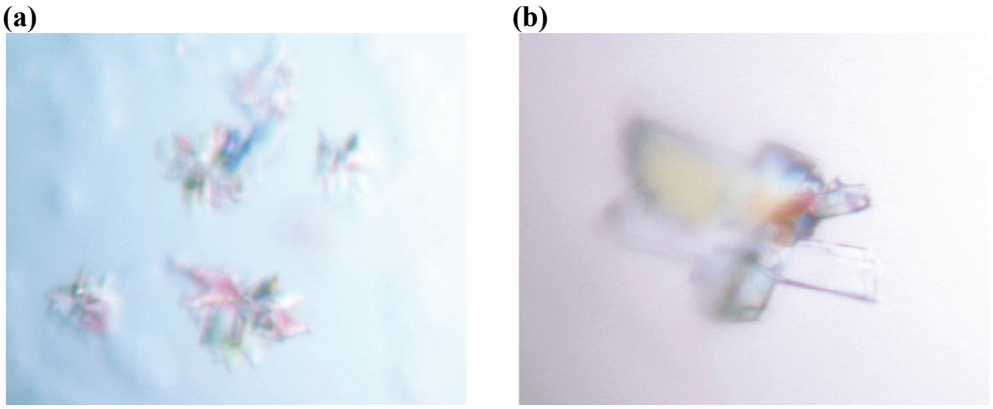


Figure 4-1 Crystallisation of 21 bp *BDNF* fragment with MBD domain
(a) spontaneous nucleation and (b) clusters from spontaneous nucleation was crushed and used as seeds for microseeding. Crystals grew after 24 hours post seeding. Fully grown crystals has maximum dimension approximately 0.5mm. These crystals diffracted X-rays to 20 and 9Å, respectively, at synchrotron sources.

As for 20 bp DNA duplex (oligos number 4; Table 4-1) in complex with construct 77-167, plate-like crystals (Figure 4-2a) were grown from Natrix screen's number 27 (Table 4-2). These crystals diffracted X-rays to 4 - 5Å on the home source. Microseeding with these crystals as seeds in a lower precipitant condition [26 – 27% (w/v) PEG8000] improved the crystal shape (Figure 4-2b) and additional seeding with

serial seed dilutions successfully grew crystals with maximum dimension 0.8mm (Figure 4-2c) which led to the native dataset (*Native*; Table 4-3). The wild type DNA-protein cocrystal diffracted X-rays (wavelength 0.9794Å) to 2.7Å resolution at station BM14, ESRF Grenoble.

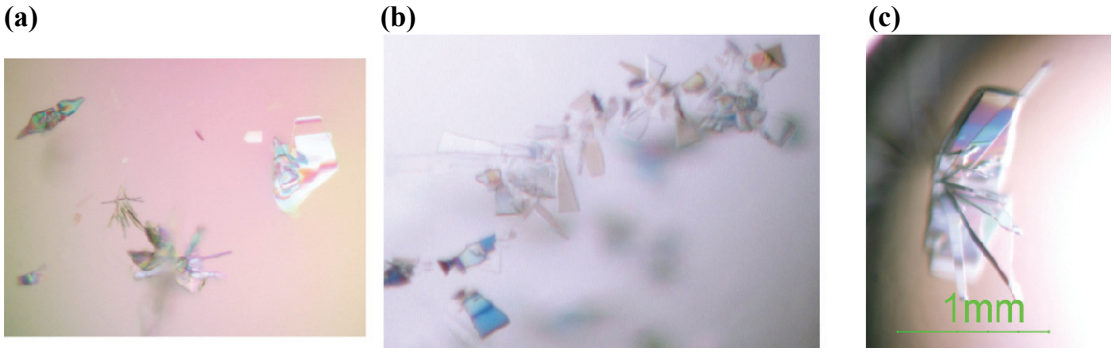


Figure 4-2 Crystallisation of 20 bp *BDNF* fragment with MeCP2 MBD domain
(a) Growth of crystals from 30% (w/v) PEG8000, 200mM Ammonium acetate, 10mM magnesium acetate and 50mM Na Cacodylate pH6.5. (b) microseeding with initial crystals (a) as seed stock. (c) Crystals grew in 27% (w/v) PEG8000 from diluted seed stock. Fully grown crystals with maximum dimension approximately 0.8mm and diffracted X-ray to 2.7Å.

4.3.1.2 SDS-PAGE analysis

In order to verify if these crystals were indeed a protein-DNA complex, several crystals were collected, washed with crystallisation solution, dissolved in SDS loading buffer, boiled and analysed on an SDS-PAGE (Figure 4-3). An aliquot (0.5µl) of drop solution was concurrently analysed in the same gel. From the Coomassie blue staining, a protein band corresponding to approximately 11kDa was detected in both preparations. Surprisingly, another species with molecular weight approximately 30kDa was also found in the melted crystal solution (Figure 4-3, lane 1). This unknown species could be covalently linked dimer of the protein or DNA. A duplicate of the gel was stained with ethidium bromide. Two bands were detected which were believed to be denatured oligonucleotides (lower band) and reannealed DNA duplexes (upper band) in the stained gel. These results strongly argued that the crystal containing both protein and DNA.

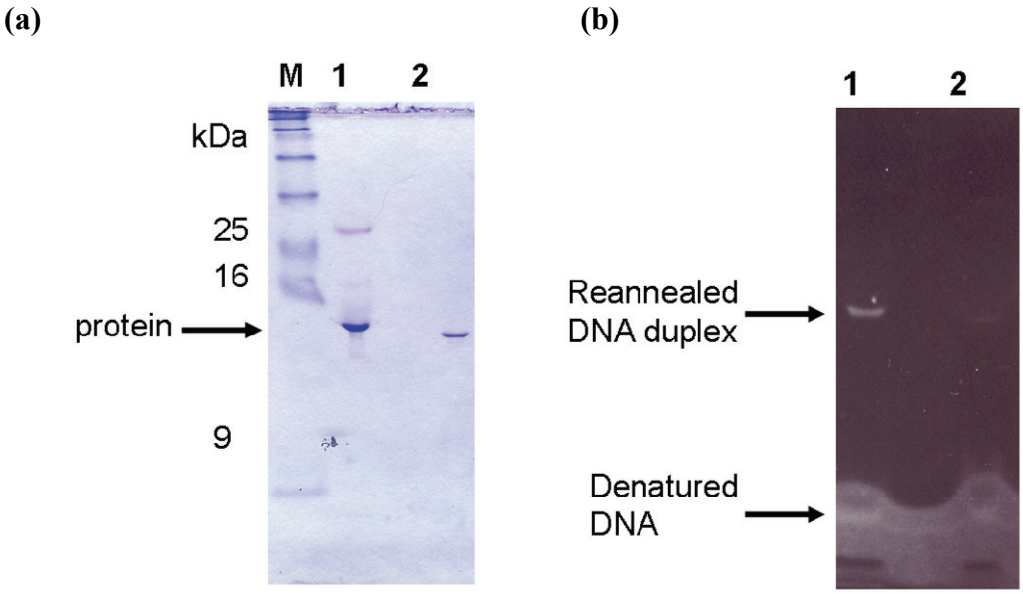


Figure 4-3: SDS-PAGE analysis of protein-DNA cocrystal.
Several crystals were mounted and dissolved in SDS-PAGE loading buffer, boiled for 10 min and separated on two 18% (w/v) SDS-PAGE gels (lanes 1 in both gels). An aliquot (0.5µl) of the drop solution was also analysed on the same gels (lane 2 in both gels). Gel (a) stained with Coomassie blue and gel (b) stained with ethidium bromide. Protein and DNA bands are indicated by arrows. Lane M; protein marker.

4.3.1.3 Iodinated derivatives

An attempt to solve the native structure of MeCP2 MBD in complex with the 20bp *BDNF* fragment using the native data was made. Molecular replacement using the NMR established models (Ohki *et al.*, 2001; Wakefield *et al.*, 1999) and various standard B-DNA as phasing models were not successful in generating an interpretable electron density map for the protein region although reasonable solutions were produced. This result indicated that experimental phasing using heavy atoms would be required to solve the X-ray structure. A detailed explanation of molecular replacement will be discussed in the experimental phasing section in this chapter.

In order to introduce a heavy atom into the DNA-protein co-crystal, 5'iodo-uracil (5'IdU) was directly incorporated into the oligonucleotides during synthesis. This was achieved by replacing T3 and T17 with 5'IdU of one strand of the duplex (oligos number 5; Table 4-1). Co-crystallisation of iodinated DNA and construct 77-167 was carried out as for native crystal. Thin plate-like crystals were grown using a similar crystallisation condition [25-27% (w/v) PEG8000, 200mM ammonium acetate, 10mM magnesium acetate and 50mM sodium cacodylate, pH6.5]. These crystals only

diffracted to 7.5Å at SRS, Daresbury (Figure 4-4a). Optimisations with various parameters, particularly PEG8000, DNA-protein complex and ion concentrations did not improve the quality of these iodinated DNA-protein cocrystals. Microseeding using the thin plate crystals as seed stock managed to increase the dimensions but not the thickness of the crystal (Figure 4-4b). Incorporation of 0.1-0.4% (w/v) of spermine, spermidine, cobalt hexamine and β-octylglucoside (Additive screens, Hampton Research) in the crystallisation solution drastically enhanced the nucleation rate and yielded relatively small crystals.

(a) **(b)** **(c)**

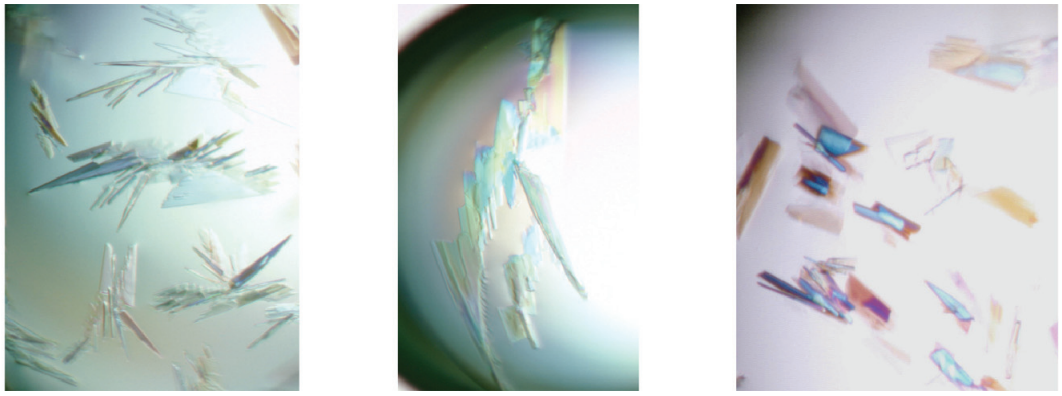


Figure 4-4 Crystallisation trials of iodo-uracil DNA-protein complex
(a) growth of crystals from 25% (w/v) PEG8000, 200mM ammonium acetate, 10mM magnesium acetate and 50mM sodium cacodylate pH6.5. (b) microseeding with spontaneous grew crystals as seed (c) a reduction in ammonium acetate concentration.

The approach followed was PEG screening. PEGs with molecular weight ranging from 350-20,000 kDa were used. Surprisingly, various PEG with molecular weight above 1000 kDa were found to be able to crystallise the iodinated DNA-protein complex even though most crystals were thin plates. Nevertheless, the crystallisation solution containing 28% (w/v) PEG2000 together with otherwise unaltered parameters; chunky crystals with maximum dimension larger than 1mm were grown within 2-3 days (Figure 4-5). The quality of the crystals including the thickness and resolution limit was greatly improved by substituting PEG8000 with PEG2000. This crystal diffracted to a maximum resolution of 3.2Å at synchrotron radiation sources. However, these crystals were metastable and dissolved within one week or upon vibration during plate observation. In subsequent experiments, crystals containing a single 5'iodo-uracil at position 3 and 17 (number 6 and 7; Table 4-1), respectively, were grown using a similar crystallisation condition.

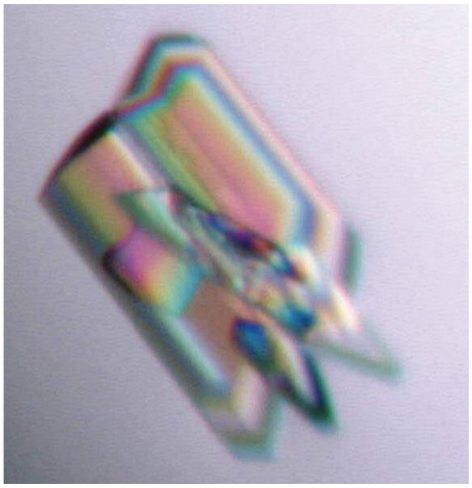


Figure 4-5: Co-crystal of 20 bp iodinated *BDNF*-MBD of MeCP2
Crystals were grown within two days in an untouched condition from 28% (w/v) PEG2000, 200mM ammonium acetate, 10mM magnesium acetate and 50mM sodium cacodylate, pH6.5. Maximum dimension of this crystal is 1mm.

4.3.1.4 Seleno-Met derivative (Wild type MeCP2)

SAD phasing using the anomalous signal from iodine located in the DNA bases generated experimental maps which also lack interpretable protein features (see experimental phasing). Improvement of the experimental map was essential to show protein features that could be interpreted for initial model building. The next strategy used was incorporation of heavy atom into the protein. One of the common approaches is introduction of seleno-Met into the recombinant protein during protein production using a Met auxotroph *E. coli* strain (Hendrickson *et al.*, 1990). The native (sulphurous) Met in the protein can be replaced efficiently with seleno-Met under carefully controlled experimental conditions. The selenium in reduced form can then be used to provide anomalous scattering at certain wavelengths which are particularly useful in experimental phasing by SAD and MAD (Hendrickson *et al.*, 1990). In this study, the first seleno-methionyl protein expressed and purified was SeMet94-MBD which contains only one Seleno-Met at position 94.

SeMet94-MBD was co-crystallised with the 20 bp *BDNF* fragment (number 4; Table 4-1) using 26-28% (w/v) PEG8000, 200mM ammonium acetate, 10mM magnesium acetate and 50mM sodium cacodylate, pH6.5 and 2mM DTT (Figure 4-6a). These rather small crystals, however, did not diffract to a resolution better than 9Å at synchrotron sources. An alternative approach was to co-crystallise SeMet94-MBD

with the DNA containing two iodine atoms (number 5; Table 4-1). This iodinated DNA-protein complex was co-crystallised using the same crystallisation condition as above and the co-crystals were grown within 3 days with the longest dimension of approximately 0.2mm (Figure 4-6b). A weak selenium signal could barely be detected in an Se fluorescence scan at SRS Daresbury. In spite of the weak Se signal, a dataset with a resolution of 2.85Å at wavelength 0.975Å (according to theoretical Se *K*-edge) was collected. In order to benefit from the iodine signal, another dataset was also collected at wavelength 2.07Å to maximise the anomalous signal. Data at the iodine adsorption edge (~2.4Å) cannot be collected because this wavelength beyond the tuneable range of SRS, Daresbury.

(a) **(b)**

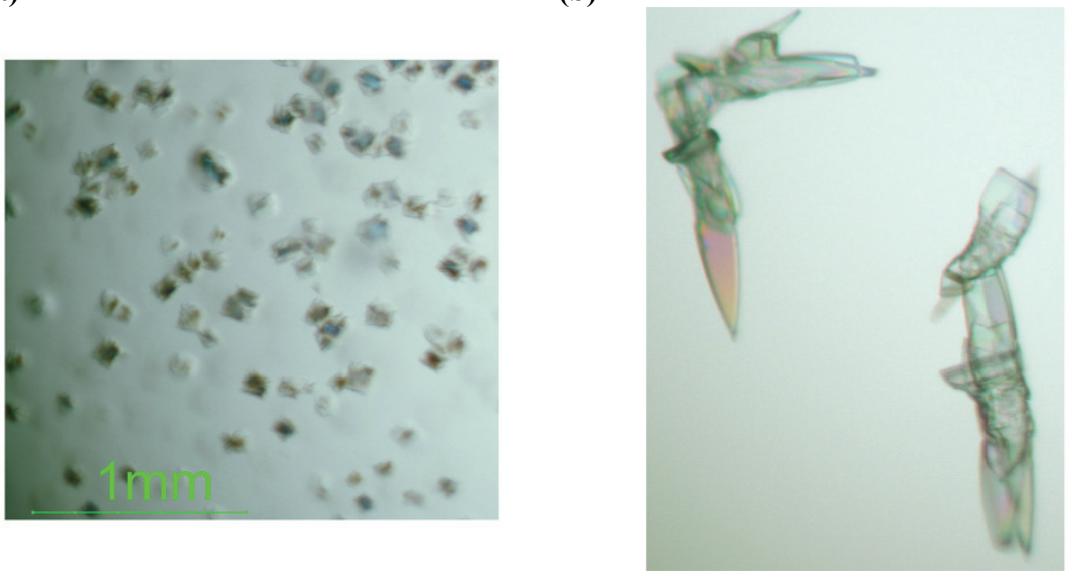


Figure 4-6: Cocrystal containing 5'iodo-uracil and seleno-Met
(a) cocrystals of SeMet94-MBD-methylated DNA complex were grown within 3 days using 27-28% (w/v) PEG8000, 200mM ammonium acetate, 10mM magnesium acetate and 50mM sodium cacodylate, pH6.5 (b) cocrystals of SeMet94-MBD-iodinated DNA with maximum dimension of 200µm under the same condition as above.

4.3.1.5 Seleno-Met derivative (mutant A140M)

The only Met (Met94) in the construct 77-167 is located at the N-terminal region of the MBD domain. As indicated in the NMR model (Wakefield *et al.*, 1999), the protein region before Asp96 is composed of mobile loops. It is possible that this mobility masked the Se anomalous signal from the crystal containing a single Se atom per 91 amino acids. A common approach to increase the number of SeMet residues in the protein is by substituting other amino acids with Met which do not significantly

alter the proteins three-dimensional structure. Following examination of the established NMR models and DNA binding analysis (Ohki *et al.*, 2001; Wakefield *et al.*, 1999), Ala140 located in the α-helix was replaced with Met using site-directed mutagenesis. The SeMet derivative of A140M (A140SeMet) was produced and crystallised with a 20 bp *BDNF* fragment (number 4, Table 4-1) using 26% (w/v) PEG2000, 200mM ammonium acetate, 10mM magnesium acetate, 50mM sodium cacodylate pH6.5 and 1mM DTT. These crystals diffracted X-rays to approximately 3Å with very streaky spots over a wide angular range (Figure 4-9a). Soaking of these crystals in the precipitant solution above but replacing magnesium acetate with manganese chloride, improved the resolution to 2.5Å. For both crystal preparations, MAD datasets were collected for experimental phasing.

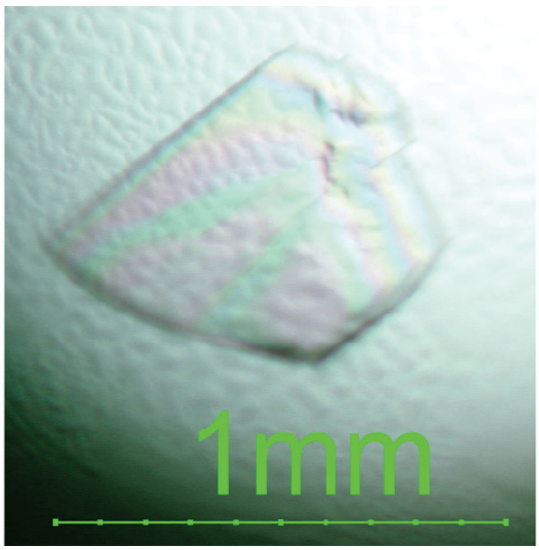


Figure 4-7 Co-crystal of A140SeMet-MBD complexed with 20bp *BDNF* fragment
This crystal was grown from precipitant solution containing 26% (w/v) PEG2000, 200mM ammonium acetate, 10mM magnesium acetate, 50mM sodium cacodylate pH6.5 and 2mM DTT at 17 ˚C. Upon soaking in solution containing 10mM manganese chloride, this crystal diffracted X-rays to a maximum resolution of 2.5Å at station BM14, ESRF Grenoble, France.

In summary, the MeCP2 MBD domain can be cocrystallised with single-base overhang 20bp methylated DNA containing symmetrical methyl-CpG dinucleotides and an adjacent AT run using 26-28% (w/v) PEG2000 or PEG8000, 200mM ammonium acetate, 10mM magnesium acetate and 50mM sodium cacodylate, pH6.5. The presence of 2mM DTT is required for cocrystallisation of seleno-Methionyl derivatives. Blunt ended DNA however did not produce reasonable size crystals for

X-ray data collection. This observation suggested that the end-to-end stacking of single-base overhangs DNA fragments (see crystal packing) is essential to form long pseudo-continuous double helices for cocrystallisation of the MeCP2 MBD domain with the 20 bp methylated DNA.

4.3.2 Data collection

Data collections were conducted mainly at Station BM14 ESRF Grenoble, France and SRS Daresbury, UK at 100K using CCD detectors. Among several datasets collected from native protein-DNA complexes and heavy atom derivatives, seven datasets will be discussed in this chapter (Table 4-3). Six out of seven datasets were collected at Station BM14 ESRF Grenoble except datasets from crystal *IodoSe* (Table 4-3), which were collected from SRS Daresbury. The native dataset was collected at wavelength 0.9794Å for total rotational angle of 180˚ with reflections out to 2.7Å resolution. A MAD dataset, extending to 2.9Å without usage of cryoprotectant was collected from A140SeMet complexed with the *BDNF* fragment. However, the dataset with the highest resolution (up to 2.5Å) was collected from a similar crystal which has been soaked in a precipitant solution containing Mn^{2+}. Additionally, the best diffraction resolution for datasets containing iodine ranged between 2.65 to 2.85Å (Table 4-3).

As part of the strategy in collecting datasets containing the anomalous scatterer selenium, all seleno-derivative crystals were scanned for normal and anomalous contributions from the incorporated selenium atoms. Figure 4-8 shows the absorption and dispersion spectra of the selenium signal from the cocrystal of A140SeMet complexed with methylated DNA. The fluorescence scan was conducted from 12.63 keV to 12.73 keV and was used to determine the wavelengths for maximum adsorption (Se peak) at 12678.23 keV (0.9780 Å) together with its *f''* (4.45) and *f'* (-7.28) values, and the inflection point (IP) at 12674.10 keV (0.9785 Å) together with its *f''* (2.15) and *f'* (-8.53) values (Table 4-3). The wavelength in the high remote area was set at 0.9080 Å. For each wavelength, 360 oscillation images with 1˚ interval were collected.

Table 4-3 Reflection data statistics for data processed in space group C2

Crystal	***Native***	***D21***	***D22***	***IodoSe***		****A140SeMet***			****A140SeMet-Mn[b]***		
						MAD1			***MAD2***		
Dataset	***Native***	***Iodo17***	***Iodo3***	***IodoSe1***	***IodoSe2***	***Peak***	***IP***	***HR***	***Peak***	***IP***	***HR***
Temperature (K)	100	100	100	100	100	100	100	100	100	100	100
Wavelength (Å)	0.9794	1.459	1.6085	0.975	2.0700	0.9780	0.9785	0.9080	0.9780	0.9785	0.9080
Space group	C2	C2	C2	C2	C2	C2	C2	C2	C2	C2	C2
Unit cell											
a (Å)	82.24	86.20	87.22	85.12	85.12	87.57	87.57	87.57	79.71	79.97	79.87
b (Å)	53.92	51.73	53.87	51.04	51.04	53.83	53.83	53.83	53.60	53.71	53.74
c (Å)	63.24	65.98	65.95	65.62	65.62	67.75	67.75	67.75	65.73	65.94	65.93
β (°)	128.17	136.57	130.17	136.72	136.72	130.78	130.78	130.78	132.10	132.11	132.18
Outer shell resolution (Å)	2.85 - 2.70	2.79 - 2.65	2.95 - 2.80	3.00 – 2.85	3.00 – 2.85	3.06 - 2.90	3.06 - 2.90	3.06 – 2.90	2.64 - 2.50	2.65 - 2.51	2.64-2.50
R_{sym} (%)[a]	7.5 (50.6)	11.5 (46.5)	8.9 (70.5)	7.2 (20.9)	8.7 (28.7)	11.8 (46.0)	11.9 (45.2)	11.6 (43.2)	7.5 (37.7)	7.7 (86.3)	9.9 (159.0)
R_{pim} (%)[a]	4.7 (35.9)	5.7 (23.8)	3.4 (27.9)	6.9 (19.5)	7.9 (24.8)	10.9 (42.1)	11.0 (41.1)	10.7 (39.6)	4.5 (28.1)	4.9 (65.0)	6.0 (95.3)
R_{meas} (%)[a]	8.9 (62.5)	12.8 (52.3)	9.0 (75.9)	10.0 (28.7)	11.8 (38.2)	16.1 (62.5)	16.2 (61.3)	15.8 (56.5)	8.8 (47.4)	9.1 (108.9)	11.6 (185.7)
Completeness (%)[a]	97.5 (87.3)	100.0 (100.0)	96.2 (94.6)	97.5 (95.6)	95.6 (91.9)	99.8 (100.1)	99.8 (100.1)	99.8 (100.1)	98.5 (90.7)	98.4 (89.6)	99.8 (100.0)
Anomalous completeness (%)[a]	-	100.0 (100.0)	96.6 (94.9)	84.8 (77.5)	70.8 (60.0)	99.7 (99.9)	99.6 (100.0)	99.7 (100.1)	98.4 (89.9)	98.0 (87.9)	99.9 (100.0)
Multiplicity [a]	3.5 (2.9)	9.4 (9.4)	14.0 (14.2)	3.5 (3.4)	3.5 (3.4)	3.9 (3.9)	3.9 (4.0)	3.9 (4.0)	7.1 (5.3)	6.4 (4.9)	7.3 (7.3)
Anomalous multiplicity	-	4.9 (4.8)	7.2 (7.2)	1.9 (1.8)	1.9 (1.9)	2.0 (2.0)	2.0 (2.0)	2.0 (2.0)	3.6 (2.7)	3.3 (2.5)	3.8 (3.8)
Unique reflections [a]	5914 (750)	5910 (859)	5616 (794)	4464 (623)	4371 (598)	5245 (775)	5247 (775)	5245 (775)	7122 (950)	7090 (940)	7178 (1045)
$\langle I \rangle / \sigma(I)$[a]	13.5 (1.6)	18.9 (3.4)	28.9 (3.1)	12.5 (4.2)	12.2 (3.3)	11.2 (2.1)	11.2 (2.1)	11.5 (2.2)	19.7 (3.0)	17.4 (1.4)	15.5 (1.0)
f'' (f')	-	-	-	-	-	-7.28 (4.45)	-8.53 (2.15)	-	5.8 (-7.11)	2.58 (-9.78)	-

[a] Value in parenthesis are for the highest resolution shell

[b] This crystal was soaked in solution containing 35% (w/v) PEG2000, 200mM ammonium acetate, 10mM manganese chloride and 50mM sodium cacodylate pH6.5 for 15 min prior to flash frozen the crystal in liquid nitrogen

* These crystals were mounted from the same crystallisation drop

The datasets from iodine derivatives (crystal *D21*, *D22* and *IodoSe*; Table 4-3) were collected in a way to maximise the anomalous signal from iodine. Because of the absorption edge wavelength of iodine (~2.4 Å) is beyond the adjustable synchrotron wavelength range, the datasets containing 2 atom of iodine per DNA molecule were collected at 2.07 Å at SRS Daresbury whereas other iodinated datasets were collected with a rather different strategy at 1.61 Å (dataset *Iodo3*, Table 4-3) and 1.46 Å (dataset *Iodo17*, Table 4-3) in a way to compromise the anomalous signal for the X-ray intensity (Table 4-3).

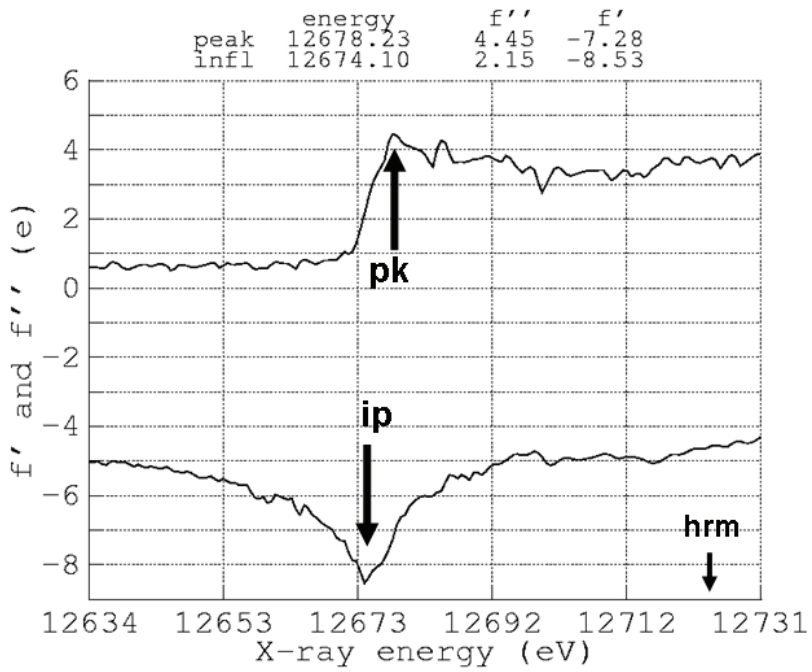


Figure 4-8 Plot of normal and anomalous scattering of selenium
The fluorescence scan was conducted from 12.63 keV to 12.73 keV and the wavelengths for maximum absorption edge or selenium peak (pk), the minimum normal scattering or inflection point (IP) were determined. The high remote (hrm) wavelength was selected manually. The *f"* and *f'* of the selenium peak and inflection point were also identified from the scan.

4.3.3 Data processing and integration

All seven datasets were auto-indexed and integrated using the MOSFLM (Leslie, 1992) package. A general procedure was applied to all datasets except specific considerations discussed here. The mosaicity estimations varied considerably (~between 0.5 to 1.3˚) at different rotational angles. Therefore, MOSFLM failed to perform cell refinement in a reasonable way. To omit this refinement step during unit cell and space group determinations, images separated at 10˚ intervals with the first and the last images separated by at least 90˚ were chosen for spot searching and auto-

indexing. The unit cell dimensions, mosaicity and tilt angles were fixed during data integration in order to process a consistent dataset. All data were reduced and scaled with SCALA and the R_{merge} value of all images were examined manually and frames with unusual high R_{merge} value were removed. The final X-ray structure of MeCP2 MBD complexed with methylated DNA was determined from a cocrystal (crystal *A140SeMet-Mn*, Table 4-3) which had been soaked in solution containing 10mM Mn acetate. Details of space group and unit cell determination are presented in the following sections.

4.3.3.1 Space group determination

The co-crystal *A140SeMet-Mn* (Table 4-3) was transferred from the mother liquor to a precipitant solution containing 35% (w/v) PEG 2000, 200mM ammonium acetate, 10mM manganese chloride and 50mM sodium cacodylate pH6.5. The crystal was soaked for 15 min before flash cooling in liquid nitrogen. This process unexpectedly improved the maximum resolution to 2.5Å with strong DNA diffraction pattern at ~3.3 Å (Figure 4-9) which might correspond to the average distances of B-DNA base-stacking. The DNA diffraction pattern in fact has been observed in all data with resolution better than 3.2Å. Auto-indexing with MOSFLM identified the most likely Bravais lattice to be C-centered monoclinic (space group C2) with unit cell dimensions of a = 79.71 Å, b = 53.60 Å, c = 65.73 Å, $\alpha = \gamma = 90$ ° and $\beta = 132.10$ °. This was further supported by unmerged data analysis using POINTLESS (Evans, 2006) which also suggested the monoclinic Laue group C 1 2/m 1, consistent with monoclinic space group C2. The selenium peak data from crystal *A140SeMet-Mn* was successfully indexed and processed to R_{sym} and R_{pim} of 7.5 % and 4.5 %, respectively. The R_{sym} values of high resolution data deteriorated rapidly from inflection (86 %) to high remote (159%) data as a result of prolonged X-ray exposure (Table 4-3). This indicates that the crystal suffered from serious radiation damage after Se peak data collection.

4.3.3.2 Unit cell contents

The molecular weight of mutant A140M and methylated DNA are 11163 Da and 12260 Da, respectively. The calculated Matthew coefficient (Matthews, 1968) using cell content calculator in CCP4 is 2.22Å^3/Da with solvent content estimation of 53 %. This clearly indicated that there is one DNA-protein complex per asymmetric unit.

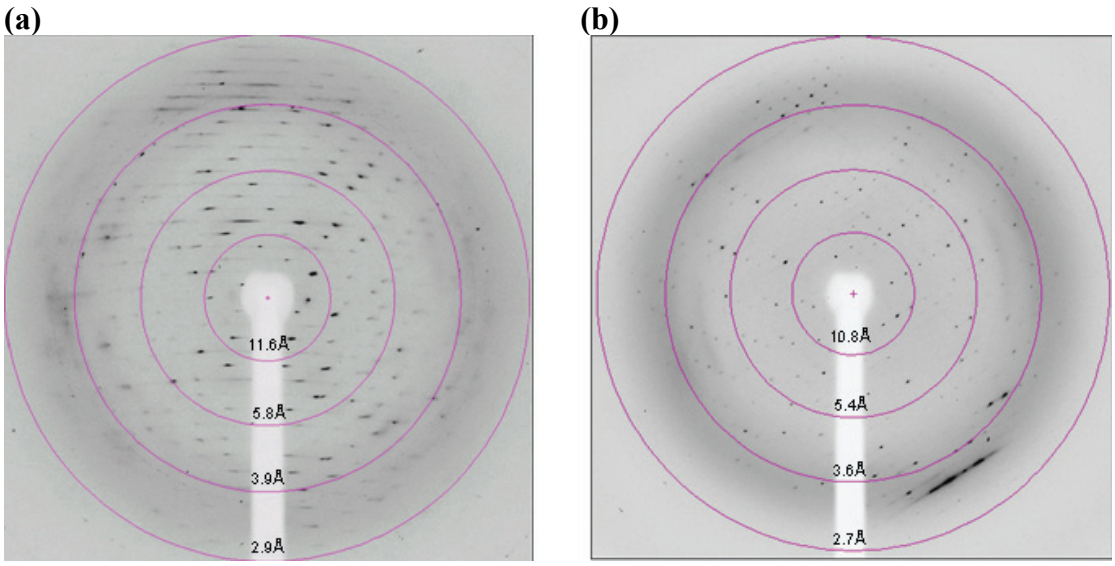


Figure 4-9 Oscillation images from crystals *A140SeMet* and *A140SeMet-Mn*
(a) Oscillation image recorded from crystal containing A140SeMet and 20 bp DNA showing that most of the spots are streaky over a wide rotational angle. This crystal diffracted to 2.9Å (b) Oscillation image recorded from a similar crystal which has been soaked in $MnCl_2$ for 15 min prior to flash frozen in liquid nitrogen. Less streaky spots were observed compared to image (a) and the maximum resolution improved to 2.5Å. Strong reflections at approximately 3.3Å might correspond to DNA base-stacking. Resolution rings are indicated.

4.3.4 Molecular phasing

The experimental phasing programmes used to solve the phase problem of the dataset collected from the cocrystal of *A140SeMet-Mn* (Table 4-3) complexed with methylated DNA were SOLVE and RESOLVE (Terwilliger, 2004) and SHELX C/D/E packages (Sheldrick, 2008). The complete structural model of the DNA-protein complex was then used as a phasing model in molecular replacement to determine the native structure and the iodinated structures (datasets *Native*, *Iodo3,* and *Iodo17*; Table 4-3). Model building and refinement were conducted iteratively with COOT (Emsley and Cowtan, 2004) and REFMAC 5.2 (Murshudov *et al.*, 1997).

4.3.5 Why experimental phasing with mutant A140SeMet?

Initially, the NMR established models of MBD1 MBD complexed with methylated DNA (Ohki *et al.*, 2001) and MeCP2 MBD alone (Wakefield *et al.*, 1999) were used as a phasing model in solving the native structure from crystal *Native* (Table 4-3) using molecular replacement in MOLREP (Collaborative, 1994) and PHASER (McCoy *et al.*, 2007) but all trials failed to yield a satisfactory solution. However,

when using various length of ideal B-DNA as a phasing model, molecular replacement successfully found the correct position of DNA; rigid body refinement led to an R_{cryst}/R_{free} of 56.3/57.9%. The resulting electron density map was heavily dominated by the double helical features of the B-DNA (Figure 4-10a). This solution was initially doubted because the protein features could not be traced. As a result, 5'iodouracil (5'IdU) was incorporated into the DNA during oligonucleotide synthesis, in which, the anomalous signal from iodine can be used for experimental phasing. In crystal *D21* and *D22*, nucleotides T17 and T3 (Table 4-3), respectively, were replaced with 5'IdU. For crystal *IodoSe*, both T3 and T17 were simultaneously replaced with 5'IdU. Both SOLVE and SHELX D managed to find the iodine positions correctly. SAD phasing with the anomalous contribution from iodine led to an electron density map which is again dominated by the DNA features. Although the electron density map still failed to display any interpretable electron density for the protein region, it confirmed the MOLREP solution above. In order to improve the electron density map, SeMet was incorporated into the wild type and later into the A140M mutant, from which the X-ray structure of the MeCP2 MBD domain in complex with the 20 bp methylated DNA was eventually determined.

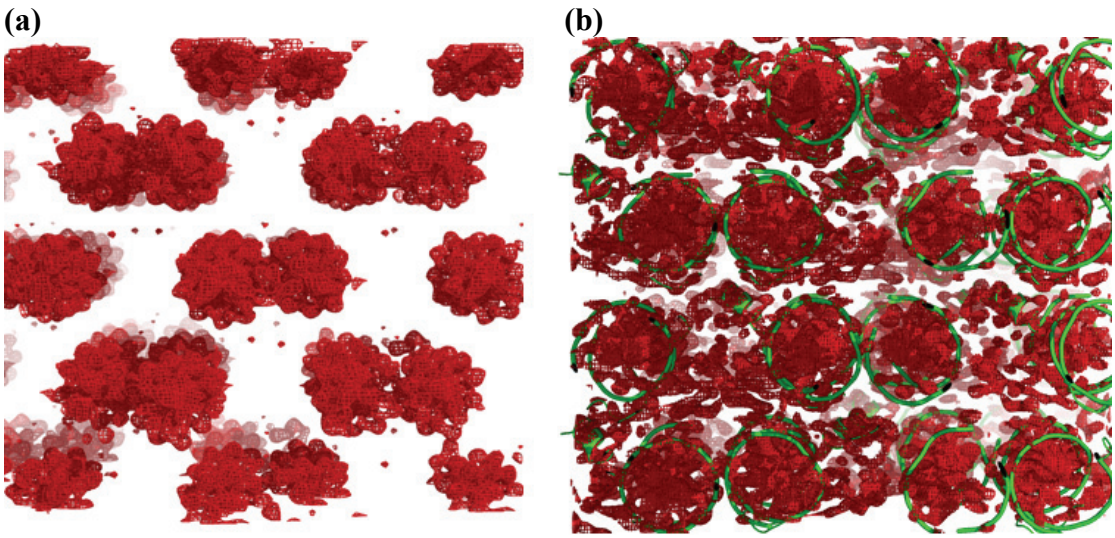


Figure 4-10 MOLREP solution with standard B-DNA
(a) Electron density map generated from molecular replacement using the 18 bp ideal B-DNA as phasing model compared with (b) Partial model of MeCP2 MBD domain complexed with methylated DNA in the first experimental map from SOLVE and RESOLVE (see SAD phasing in later section).

The wild type MeCP2 MBD domain contains only one Met (Met94), which is located at the N-terminus of the MBD domain according to Wakefield *et al.* (1999). The cocrystals of selenomethionine substituted protein and methylated DNA had been grown but failed to display a significant Se anomalous signal from an X-ray dataset collected with the wavelength set at the selenium *K*-edge. The weak anomalous signal in the X-ray data might because of an insufficient number of Se atom in the protein-DNA complex. In practice, the generally accepted minimum is one Se atom for every 14 kDa protein (Sharff *et al.*, 2000). Therefore, one single Se atom in the MeCP2 MBD complexed with a methylated DNA (total size approximately 23 kDa) may not provide sufficient anomalous signal to solve the phase problem. To enhance Se anomalous scattering power, Ala140 was mutated to Met by site directed mutagenesis. The resulting mutant; A140M; contains two Met residues (Met94 and Met140). It has been shown that Ala140 is located in the only α-helical region of the MeCP2 MBD domain (Wakefield *et al.*, 1999). The seleno-methionyl protein preparation should place an additional 'structured' Se atom in the α-helical region of the protein. Together with SeMet94, SeMet140 could potentially contribute an anomalous signal for estimating phases in methods such as SAD and MAD.

4.3.6 Multi-wavelength anomalous dispersion (MAD)

The first MAD dataset was collected from a crystal containing A140SeMet (dataset *MAD1*; Table 4-3) without using a cryoprotectant. This crystal diffracted X-rays around Se absorption edge wavelengths to a maximum resolution of ~3.0 Å. Image examination revealed that most of the frames displayed streaky spots which substantially interfered with data processing and autoindexing. MAD or SAD phasing using various programmes such as SHELX C/D/E (Sheldrick, 2008), SOLVE and RESOLVE (Terwilliger, 2004) or AutoSHARP (Vonrhein *et al.*, 2006) have been extensively tried but failed to generate a correct set of phases.

Mg^{2+} is an essential component to enable crystallisation although it is not required for the formation of MeCP2-methylated DNA complexes. However, Mg^{2+} cannot display anomalous scattering that is required for MAD or SAD. In contrast, Mn^{2+} with atomic mass of 54.9 would produce anomalous scattering at its absorption edge around wavelength 1.896Å. Therefore, substitution of Mg^{2+} with Mn^{2+} would potentially

provide an anomalous signal in determining the phases of the heavy atom substructure. Attempts to cocrystallise the construct A140SeMet complexed with methylated DNA using precipitant solutions containing Mn^{2+} has not succeeded in yielding any co-crystal of MBD-DNA complex. As an alternative choice, the co-crystal was soaked in a modified precipitant solution containing 35% (w/v) PEG2000, 200mM ammonium acetate, 10mM $MnCl_2$ and 50mM Na cacodylate pH6.5 for 15 min prior to flash freezing in liquid nitrogen. Surprisingly, some significant changes were observed from Mn soaked crystal (crystal *A140SeMet-Mn*, Table 4-3) compared to un-soaked ones, these changes include:

(i) improved maximum resolution from ~3.0 Å to ~2.5 Å,

(ii) the interference of streaky spot were greatly reduced, and

(iii) most intriguingly, changes in unit cell dimensions (Table 4-3). As a result, data collected from crystal *A140SeMet-Mn* was used for experimental phasing which eventually led to the final structural determination.

Analysis for significant anomalous scattering of Se in the cocrystal *A140SeMet-Mn* was performed mainly using two programmes; SHELXC/D/E packages (Sheldrick, 2008) and PHENIX (Adams *et al.*, 2002). Three- or two-wavelength MAD was not successful in solving the phase problem. This was probably due to the deteriorated quality of the inflection and high energy remote data. As part of the strategy to maximise the reflection intensity, one minute exposure time was allowed per oscillation image per 1˚ rotational angle. This approach obviously accelerated the crystal deterioration because of longer X-ray exposures, as indicated by the high R_{sym}, R_{pim} and $\langle I \rangle/\sigma(I)$ values for the inflection (IP) and high remote (HR) data (crystal *A140SeMet-Mn*, Table 4-3). After peak data were collected, the outermost R_{sym} increased from 37.7 % (Se peak) to 86.3 % for inflection data and this was further exacerbated for data collected at high remote area. The outermost $\langle I \rangle/\sigma(I)$ value was decreased from 3.7 (peak) to 1.4 (inflection) and 1.0 (high remote). Nevertheless, the quality of the peak data, which diffracted X-rays to highest resolution of 2.5Å, was considered moderately good quality with an overall completeness of 98.5%, outermost R_{sym} of 37.7 %, $\langle I \rangle/\sigma(I)$ of 3.0 and anomalous multiplicity of 2.7 (Table 4-3). Thus, the X-ray structure of MeCP2 MBD domain complexed with 20 bp *BDNF* fragment containing a central methy-CpG dinucleotides and 4 bp of A/T bases was successfully solved by SAD using the selenium peak data of the *A140SeMet-Mn* crystal.

4.3.7 SAD phasing with PHENIX – the successful case

Python-based **H**ierarchical **EN**vironment for **I**ntegrated **X**tallography (PHENIX) is a software suite for a highly automated macromolecular determination up to the stage of partial model building provided moderate resolution data are available (Adams *et al.*, 2002). In PHENIX, decision making strategies (the wizards in the programme) are made by considering all of the available information at each step in the process. The four available wizards in PHENIX are: structural solution using AutoSol, molecular replacement using AutoMR, model building or rebuilding using AutoBuild and ligands fitting using AutoFit (http://www.phenix-online.org/). All experimental phasing methods such as SIR, MIR, SIRAS, MIRAS, SAD, MAD are performed using AutoSol in PHENIX. In general, the AutoSol wizard uses the following steps in solving a macromolecular structure: (i) data input, analysis and scaling, (ii) heavy atom searching and scoring, (iii) phasing and density modification and, (iv) preliminary model building and refinement.

4.3.7.1 Data input, analysis and scaling

In PHENIX, analysis of significant anomalous contribution and heavy atom location are performed using SOLVE (Terwilliger, 2004). Using selenium peak data from crystal *A140SeMet-Mn* (Table 4-3), the anomalous signal-to-noise ratio ($|F^+-F^-|/\sigma|F^+-F^-|$) of 10 resolution bins was calculated; a ratio above 1.2 was regarded as a significant anomalous contribution (Table 4-4). Therefore, resolution above 3Å was discarded from heavy atom searching in this data. Two heavy atom sites were found in the asymmetric unit for the Se peak data with occupancy of 100 % and 86 %, respectively (Table 4-5).

Table 4-4 Statistics of the anomalous signal-to-noise ratio and data completeness against resolution

Resolution (A)	**∞-5.00**	**5.00-3.75**	**3.75-3.50**	**3.50-3.31**	**3.31-3.12**	**3.12-3.00**	**3.00-2.88**	**2.88-2.75**	**2.75-2.63**	**2.63-2.50**
% complete	98.6	99.9	99.8	100.0	99.8	100.0	100	100.0	99.6	89.6
N obs	796	1144	446	448	549	442	544	642	744	824
$S/\|F^+-F^-\|$	20.06	11.98	9.22	8.52	7.94	6.85	5.89	5.71	3.99	0.00
$N/\sigma\| F^+-F^-\|$	3.57	3.38	4.00	4.46	5.12	5.71	6.02	6.52	7.08	7.85
S/N ratio	5.61	3.54	2.30	1.91	1.55	1.20	0.98	0.88	0.56	0.00

The resolution cut-off for heavy atom search in this data was at 3Å. Abbreviations S and N refer to signal and noise, respectively.

Table 4-5 Heavy atom sites

Site	atom	Occupancy	X	y	z	B factor
1	Se	1.0235	0.8375	0.3649	0.1471	59.73
2	Se	0.8609	0.9711	0.0242	0.0524	60.00

4.3.7.2 SAD phasing and density modification

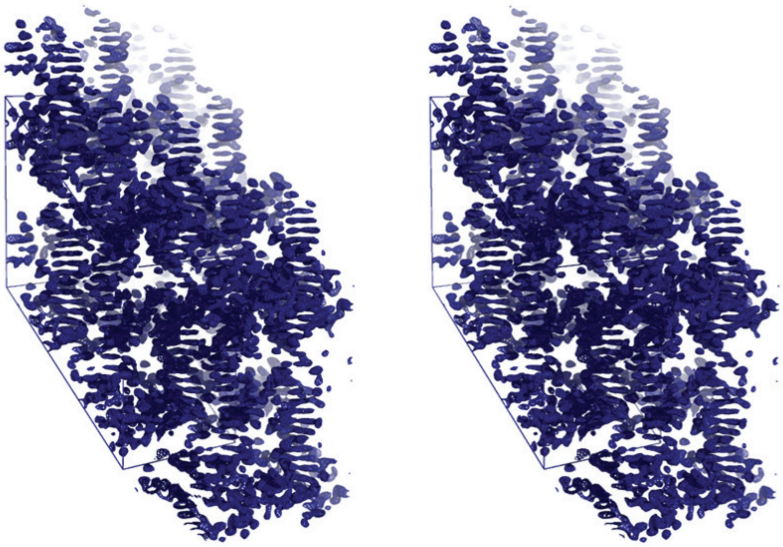
The heavy atoms found in SOLVE (SAD phasing) were refined with origin-removed Patterson refinement (Terwilliger and Eisenberg, 1983), which resulted in an electron density map with mean Figure of Merit (FOM) value of 25% (Table 4-6). Density modification using RESOLVE with a solvent content of 54% produced a final electron density map with an overall correlation coefficient (CC) and FOM of 68% and 53%, respectively. Figure 4-11a shows the first experimental map with readily interpretable double stranded B-DNA and some protein features. A closer view (Figure 4-11b) of the non-DNA electron density shows the α-helical secondary region of the protein and other continuous electron densities in the major groove of the DNA. The first selenium atom (with occupancy 1) was found to be located in the expected α-helical region (Figure 4-11b) and the second Se site is located in a relatively disordered region of the protein.

Table 4-6 FOM with resolution after experimental phasing with SOLVE

Resolution (A)	∞-9.13	9.13-5.72	5.72-4.46	4.46-3.77	3.77-3.33	3.33-3.01	3.01-2.77	2.77-2.58	**Total**
N data	330	590	759	905	1004	1129	1190	1208	**7115**
<FOM>	0.44	0.50	0.45	0.38	0.29	0.19	0.11	0.06	**0.25**

FOM can be defined as the weighted mean of the cosine of the phase error

(a)



(b)

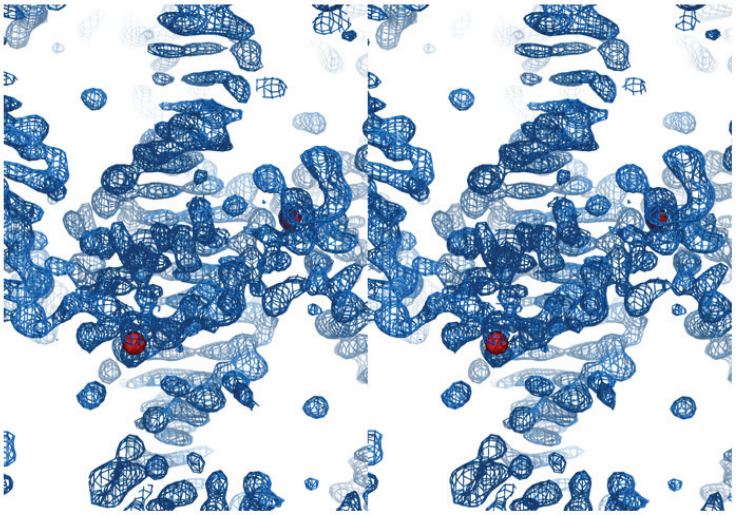


Figure 4-11 First experimental map after SAD phasing with SOLVE and density modification with RESOLVE
(a) Stereo diagram of the B-form DNA double helix and some of the protein features are shown in the first electron density map. (b) A closer stereo diagram of the only α-helical coil and other protein chains close to the major groove of the DNA are clearly seen. The two selenium atoms are represented by red spheres.

4.3.7.3 Preliminary model building and refinement

The electron density map resulting from SAD phasing was sufficient to begin an automated model building. In the PHENIX software suite, the initial macromolecular

model can be built using RESOLVE (Terwilliger, 2004). The protein and DNA partial models were built in two separate runs of model building. Thirty out of 40 bases (75 %) was built whereas only the main protein features such as α-helix and the long loop can be traced (Figure 4-12b and c). The partial coordinates of DNA and protein were combined manually and the wrongly built polypeptides in the DNA electron density were removed. The partially built model with R_{cyst}/R_{free} of 48/51% (Figure 4-12a) served as a starting model for continuous manual building. Manual building and refinement was continued iteratively using COOT (Emsley and Cowtan, 2004) and REFMAC 5.2 (Murshudov *et al.*, 1997) until the model was completed with all observable amino acids and DNA bases, the R_{cyst}/R_{free} improved to 25.8/30.1%. Model building and refinement were completed with generous helps and contributions from Dr. Iain McNae.

At this stage, the model was submitted to the TLSMD server (Painter and Merritt, 2006) for TLS motion groups determination. The server partitions the input structural model into multiple continuous chain segments based on input thermal parameters, each segment acting as a rigid group undergoing TLS (translation/libration/rotation) motion (Painter and Merritt, 2006). The multi-group TLS models were further refined with REFMAC5.2 (Murshudov *et al.*, 1997). By comparing the refinement statistics (such as FOM, RMSD and R_{cryst}/R_{free} values), the polypeptide chain was best divided into 4 TLS motion groups and each DNA strand into two TLS groups (Figure 4-13) and the average B-factor of each TLS segment are presented in Table 4-7.

The resulting model after REFMAC refinement was used to search for water molecules with COOT (Emsley and Cowtan, 2004). Using $|\mathbf{F_o}-\mathbf{F_c}|$ maps with a σ level of 2.8, COOT located 47 water molecules. The final model of the MeCP2 MBD domain in complex with 20 bp promoter region of *BDNF* gene shows good geometry with final R_{cyst}/R_{free} of 21.1/27.6 % and RMSD bonds/angles of 0.009Å /1.85°. The quality of the model was validated using PROCHECK (Laskowski *et al.*, 1993). The Ramachandran plot shows that 91.4% of residues lie within the most favoured region, only 8.6% within the additional allowed regions and none within the generously allowed or disallowed regions (Figure 4-14).

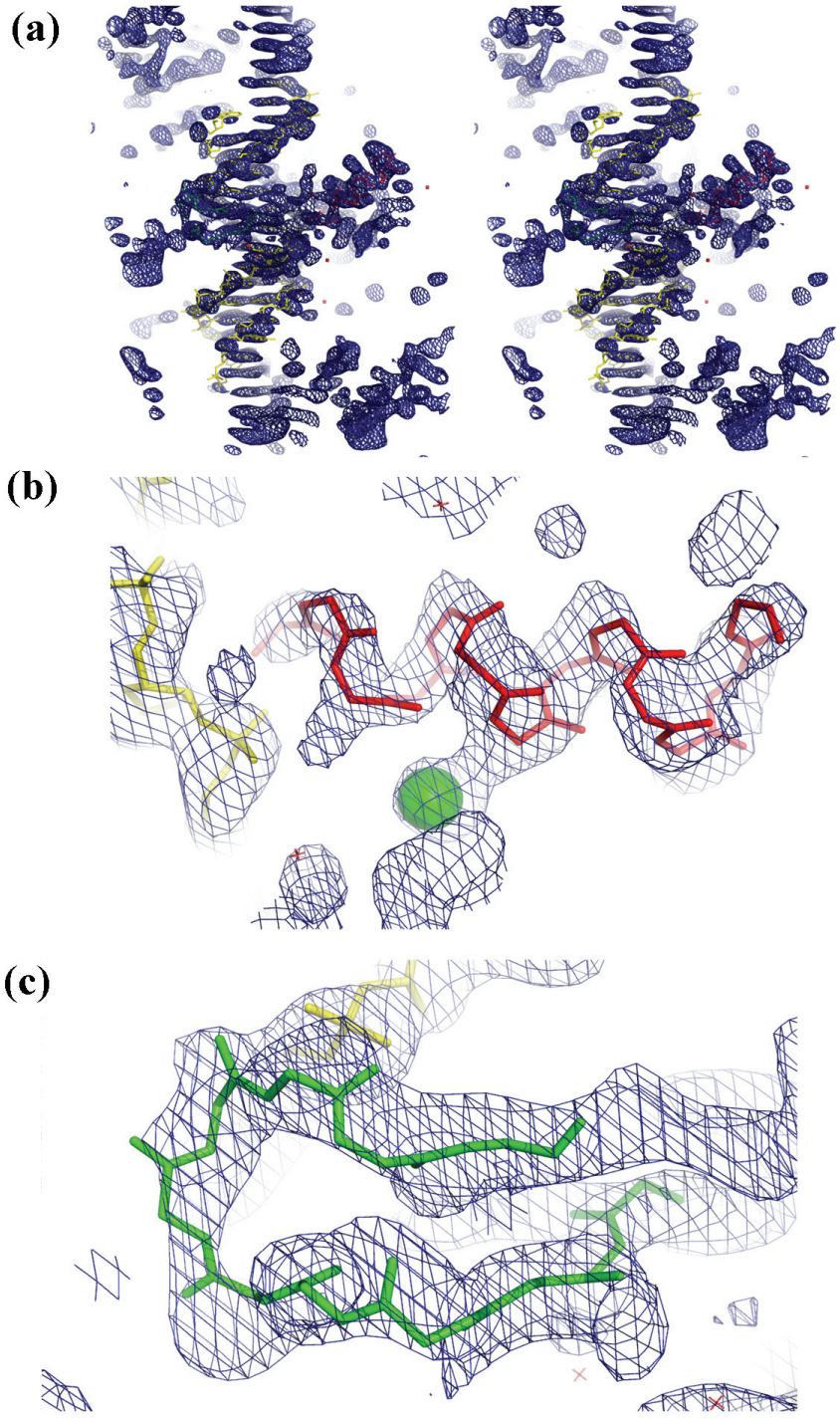


Figure 4-12 Partial model in the experimental map (automated built)
(a) stereo view of the partial model of the DNA-protein complex built with RESOLVE
(b) alpha-helical region of the protein with the selenium atom coloured in green
(c) Loop L1 of the MeCP2 MBD domain is clearly visible.

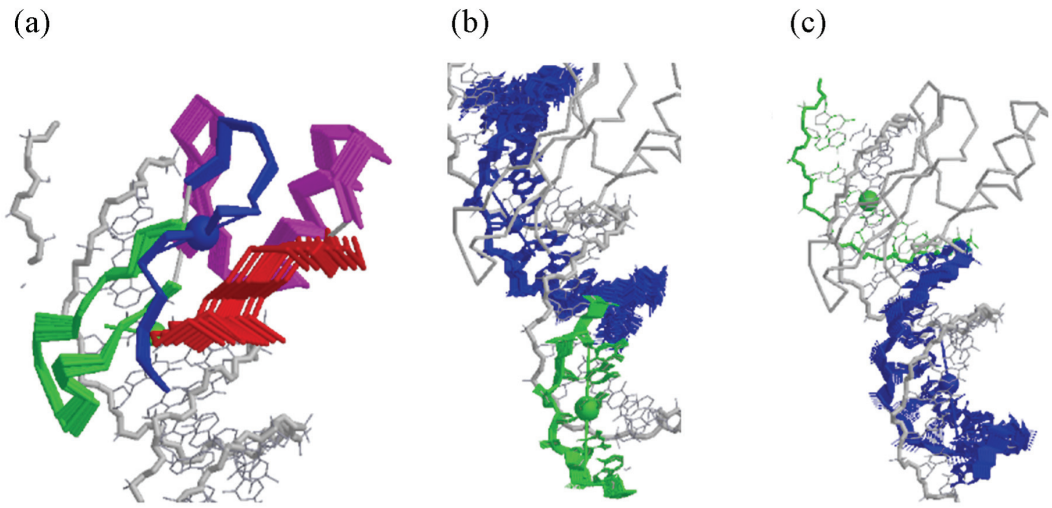


Figure 4-13 TLS motion groups of *A140SeMet-Mn* model
Model showing the TLS motions of (a) the polypeptide, (b) DNA strand (chain B) and (c) DNA strand (chain c). All TLS motion groups are coloured differently (diagrams generated using http://skuld.bmsc.washington.edu/~tlsmd/)

Table 4-7 Mean B factor of TLS motion groups

TLS segment	Number of residue	Number of Atom	Average B factor <B>/Å^2
Protein (Chain A)			
91-105	15	121	65.8
106-122	17	138	38.4
123-151	29	236	52.1
152-162	11	89	49.6
DNA (chain B)			
1-15	15	307	45.7
16-20	5	100	59.7
DNA (chain C)			
21-32	12	246	50.8
33-40	8	163	48.2

4.3.7.4 Crystal packing

Each unit cell (a = 79.71Å, b = 53.60 Å, c = 65.73 Å and β = 132.10˚) contains 4 asymmetric units (Figure 4-15). The DNA molecules stacked end-to-end to form long pseudo-continuous double helices along the c-axis. The length of each individual DNA molecule equals to the length of the c-axis. Although the individual DNA molecules are slightly bent, the long DNA double helices are absolutely straight (Figure 4-15b). Figure 4-15c shows that the DNA duplex is wrapped by two extended arms at one side of the DNA molecule. On the opposite side of the DNA molecule, two DNA molecules come into contact crystallographically but run in opposite directions. No protein-protein interaction was observed in the crystal lattice.

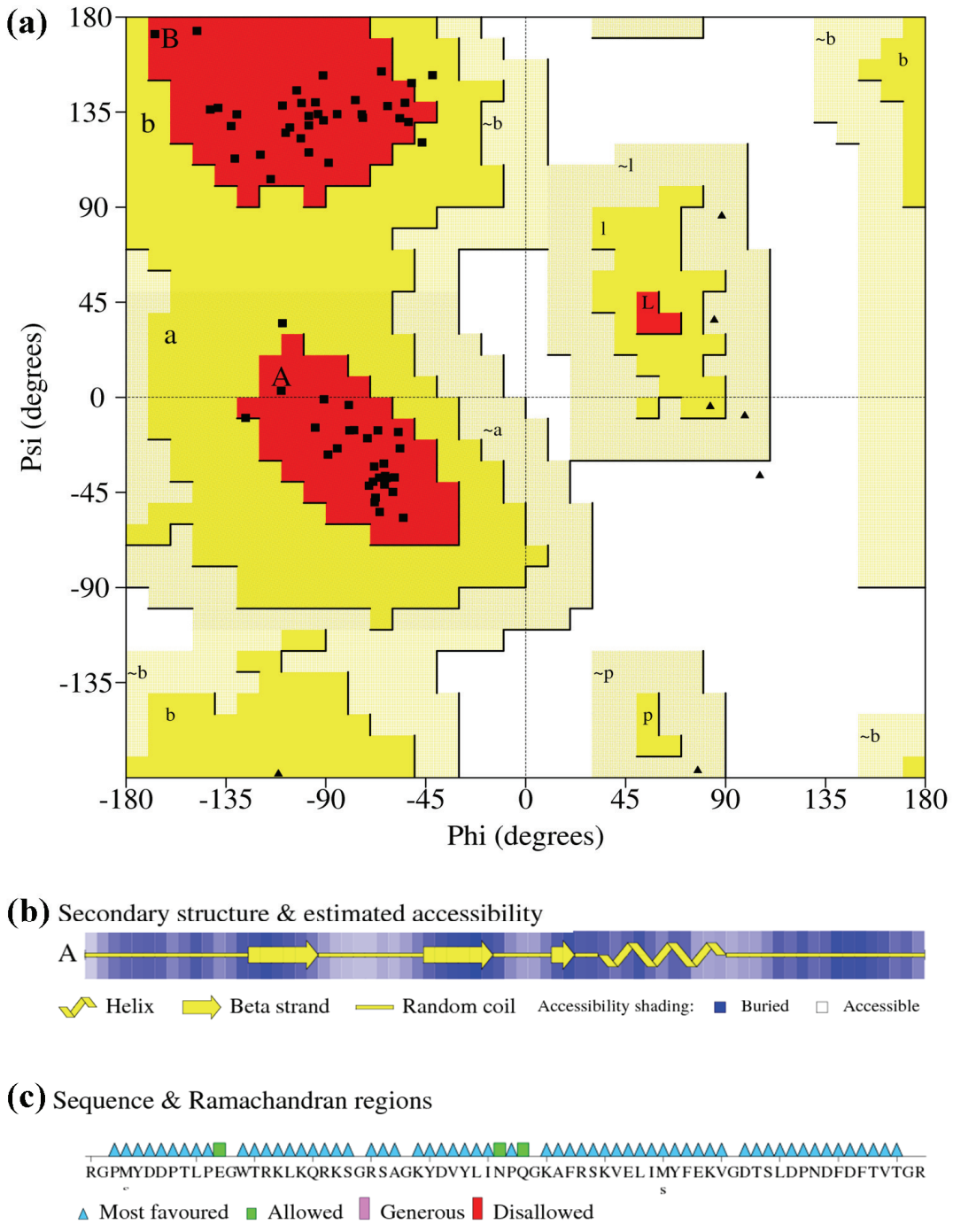


Figure 4-14 Analysis of stereochemical properties of A140SeMet-Mn
(a) Ramachandran plot shows that 91.4 % of residues located within the most favoured region, 8.6 % of residues in the additional allowed region. (b) protein secondary structure prediction and (c) amino acid sequence corresponds to Ramachandran plot.

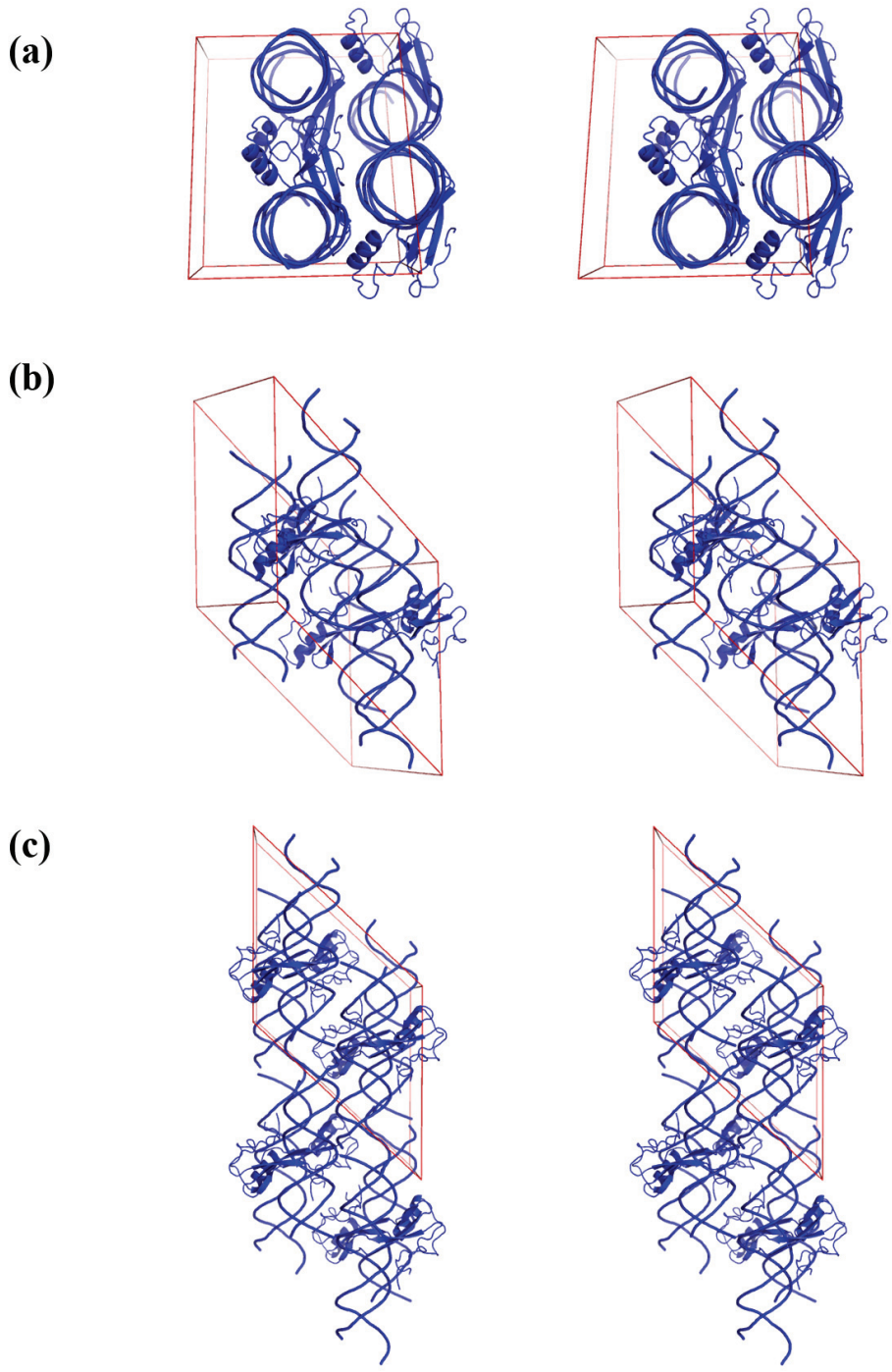


Figure 4-15 Crystal packing
(a) Stereo view along the c-axis (b) unit cell contains 4 asymmetric units (c) DNA molecules stacked end-to-end forming pseudo-continuous double helices along c-axis.

4.3.7.5 Native structure determination with molecular replacement

As discussed in section 4.3.5, molecular replacement with various lengths of standard B form DNA yielded an electron density map displaying double helical DNA features but failed to display protein features. In order to determine the MBD domain in the native structure, the final model of *A140SeMet-Mn* complexed with methylated DNA was used as a phasing model in MOLREP (CCP4, v6.0.2) (Potterton *et al.*, 2002). An initial round of rigid body refinement was carried out after molecular replacement and gave a model with R_{cryst}/R_{free} of 39.6/43.6%. The seleno-Met residues were mutated to sulphurous-Met and autofitted using COOT into the $\left|2|F_o| - |F_c|\right|$ map and this was followed by restrained refinement with tight stereochemical restraints (weighting term of 0.01 to 0.05˚) until convergence with an R_{cryst}/R_{free} of 25.1/34.9%.

At this stage, the model was submitted to the TLSMD server (Painter and Merritt, 2006) in order to perform TLS group determination. In this case, the *Native* model was segmented into 5 TLS motion groups with the polypeptide chain as one group and each DNA strand divided into 2 groups. The PDBIN and TLSOUT from the TLSMD server were used for restrained refinement in REFMAC 5.2 with tight geometry restraints. This round of refinement pronouncedly improved the convergence of the final model to R_{cryst}/R_{free} of 23.0/29.5%.

A final validation was carried out using PROCHECK (Laskowski *et al.*, 1993). The Ramachandran plot shows that 77.6 % residues within the most favoured regions, 20.7% located in additional allowed regions and 1 % in generously allowed regions. None of the residues are within disallowed regions (Figure 4-16). Nevertheless, some secondary elements (such as strand β3) could not be accurately identified.

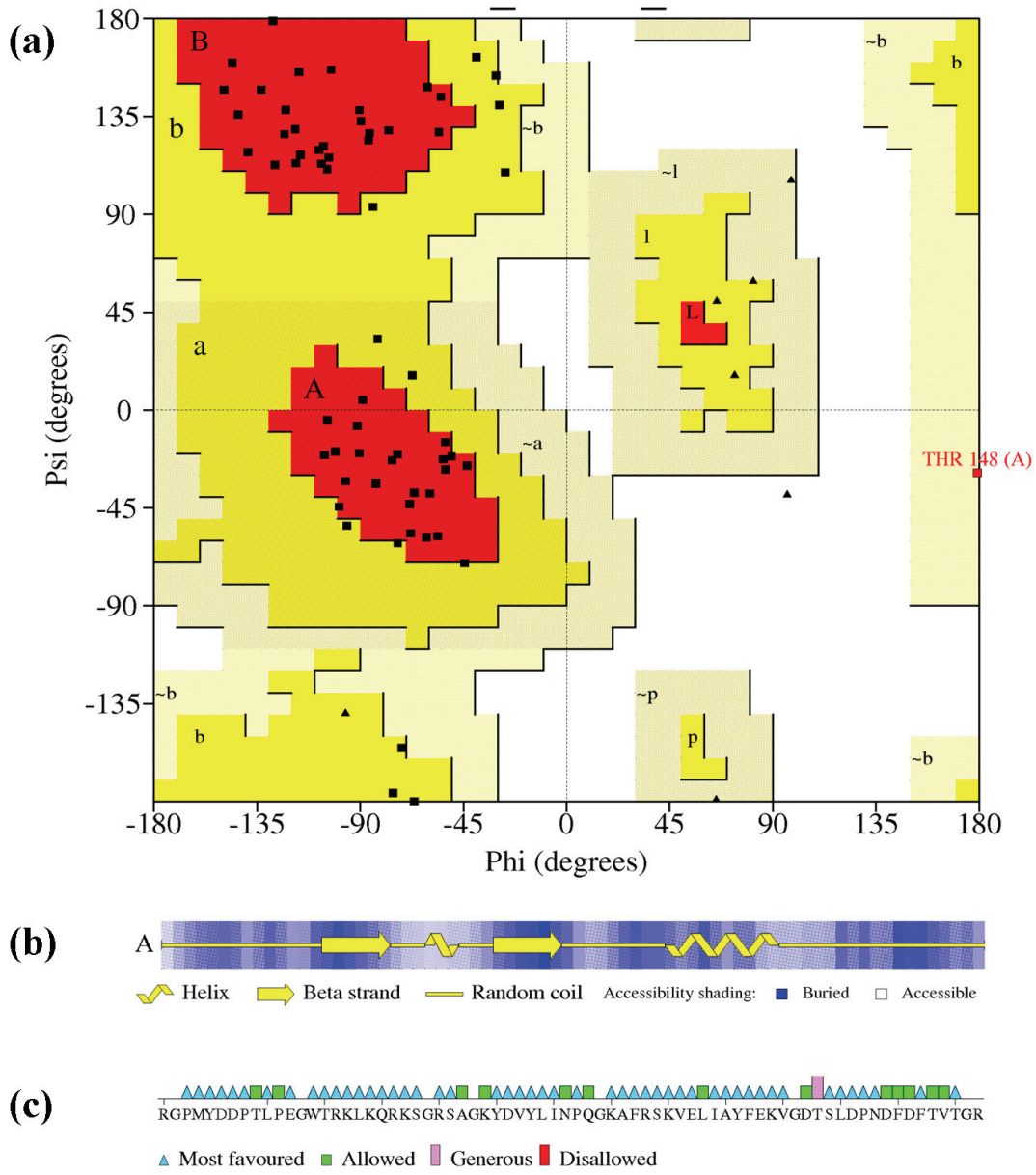


Figure 4-16 Analysis of stereochemical properties of the *Native* X-ray structure
(a) The Ramachandran plot shows that 77.6% of residues are within most favoured region, 20.7% within additional allowed regions and 1.7% located within generously allowed regions. (b) secondary structure prediction and (c) amino acid sequence corresponding to Ramachandran plot.

4.3.7.6 Iodinated structure determination using molecular replacement

As mentioned in the earlier section, the anomalous signal from the iodine successfully generated an electron density map showing the DNA double helical features. However, the iodinated data sets provide insufficient information for determining the complete structure of the MBD domain of MeCP2 complexed with a 20 mer DNA. Nevertheless, these iodinated DNA-protein complex X-ray structures (datasets *Iodo3* and *Iodo17*; Table 4-3) were eventually solved with molecular replacement (MOLREP) using the fully refined model (*A140SeMet-Mn*) as the phasing model.

For dataset *Iodo17* (Table 4-3), molecular replacement successfully found the correct position. An initial round of rigid body refinement rendered the model with R_{cryst}/R_{free} values of 52.6/53.6%. Subsequent refinement with 20 cycles of restrained refinement yielded a model with R_{cryst}/R_{free} of 39.0/49.4%. After fixing all residues and nucleotides to its original sequence in the crystal, restrained refinement produced a model with R_{cryst}/R_{free} of 25.6/31.7%. A total of 30 water molecules have been added with COOT (Emsley and Cowtan, 2004) at a sigma level 3.0 using $|F_o|$-$|F_c|$ map. The final round of restrained refinement yielded a model with R_{cryst}/R_{free} of 24.3/30.1% (Table 4-8). PROCHECK (Laskowski *et al.*, 1993) has been carried out to validate to final model. The Ramachandran plot (Figure 4-17a) indicates that 94.8% of the amino acids are within the most favoured regions, 5.2% located within additional allowed regions.

Similarly, dataset *Iodo3* (Table 4-3) was solved using molecular replacement. Initial rounds of rigid body refinement generated a model with R_{cryst}/R_{free} of 41.1/38.9%. Subsequent rounds of restrained refinement with tight geometry improved the R_{cryst}/R_{free} values to 26.3/31.4%. Further rounds of restrained refinement after fixing all the nucleotides and bases yielded a model with R_{cryst}/R_{free} of 26.6/31.8%. A total of 12 water molecules have been added using COOT (Emsley and Cowtan, 2004). The final round of restrained refinement produced a model with R_{cryst}/R_{free} of 24.9/29.5% (Table 4-8). The Ramachandran plot (Figure 4-17b) shows that 79.3% of residues are located within the most favoured regions and 20.7% amino acids are located within additional allowed regions. None of the residues has been found located in the generously allowed and disallowed regions.

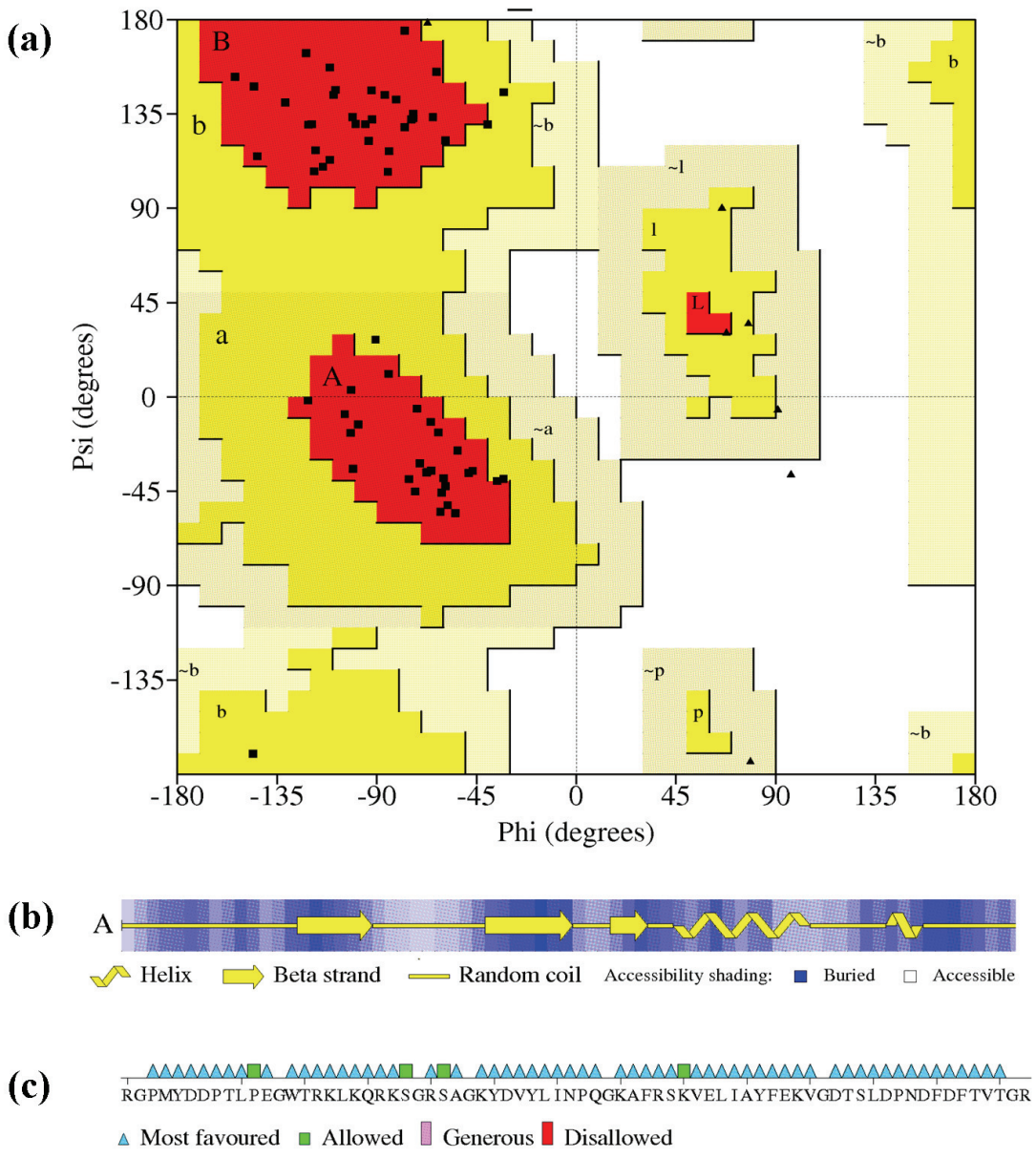


Figure 4-17 Analysis of stereochemical properties of *Iodo17* X-ray structure
(a) Ramachandran plot shows that 94.8% and 5.2% of the amino acids located within the most favoured and additional allowed regions, respectively. (b) protein secondary structure prediction and (c) amino acid sequence corresponds to Ramachandran plot.

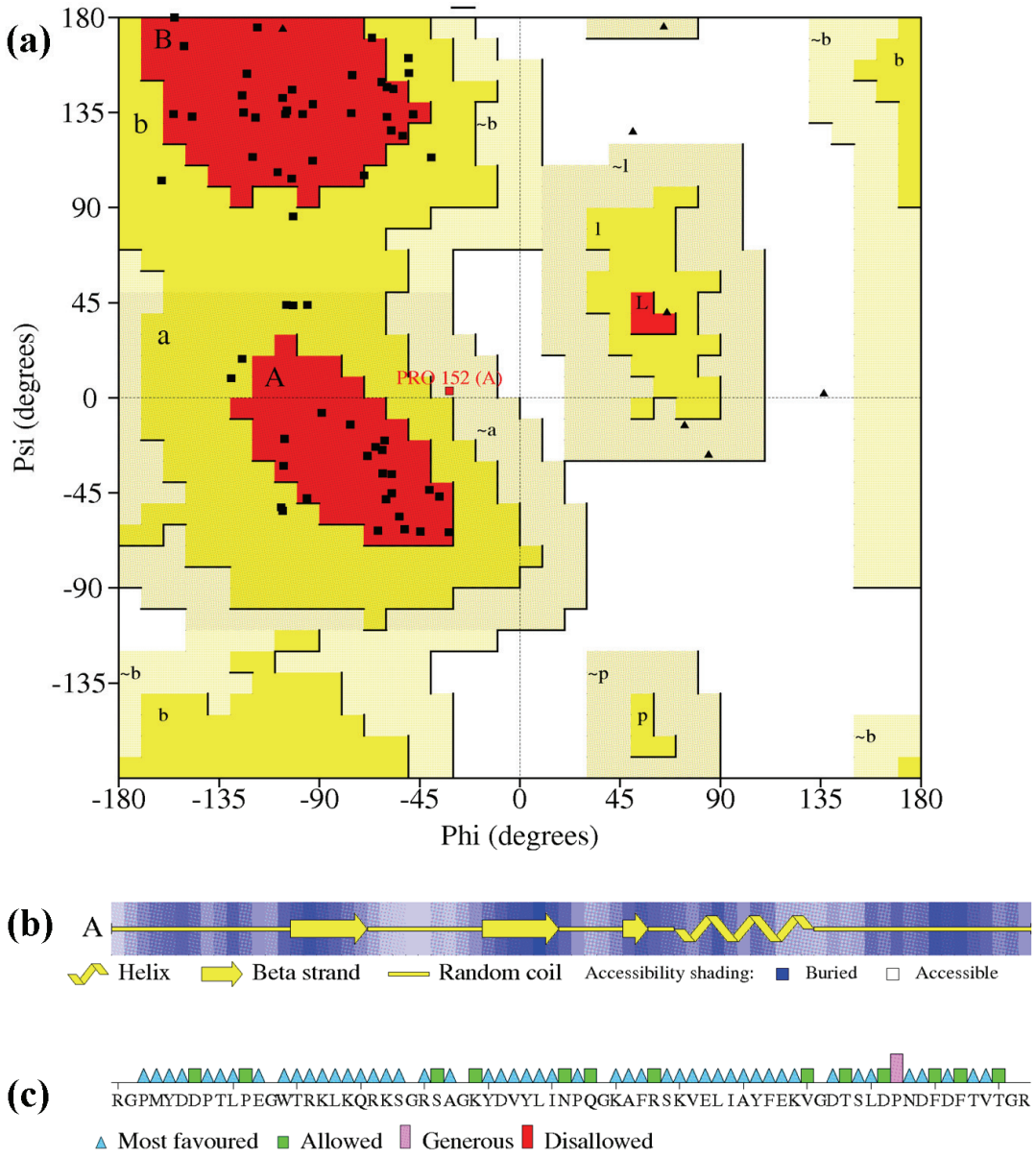


Figure 4-18 Analysis of stereochemical properties *Iodo3* X-ray structure
(a) Ramachandran plot shows that 79.3% and 20.7% of the amino acids located within the most favoured and additional allowed regions, respectively. None of the residues of these datasets located within the generous allowed and disallowed regions. (b) Protein secondary structure prediction and (c) amino acid sequence corresponds to Ramachandran plot.

4.3.8 Which is the best model?

The main objective of this study is to examine the molecular details of methyl-CpG recognition by the MBD domain of MeCP2. By comparing the quality of the 4 models that have been refined, the X-ray structures of *A140SeMet-Mn*, *Iodo3* and *Iodo17* (Figure 4-19) provide sufficient molecular information for detailed structural analysis. Table 4-8 shows that the best refined model is *A140SeMet-Mn* with a final R_{cyst}/R_{free} of 21.2/27.6 and RMSD bond/angles of 0.009/1.85. Nevertheless, the data quality of *Iodo17* is generally better than *A140SeMet-Mn* due to 100% completeness and 9.4 multiplicity. Secondary structure analysis using PROCHECK (Laskowski *et al.*, 1993) identified an additional short α-helix of ^{152}PND^{154} within MeCP2 MBD domain (Figure 4-19d). The highest redundancy data is dataset *Iodo3* which was collected from more than 600 images. The polypeptide chain of *Iodo3* displays an identical secondary structure as *A140SeMet-Mn* (Figure 4-19a, c). Overlay of iodinated models with the *Native* and *A140SeMet-Mn* indicates that the DNA-protein binding and crystal packing were not affected by the presence of iodine atom in the DNA (Figure 4-20). Out of the 4 DNA duplexes, the double helix in *Iodo17* is most deformed (Figure 4-20) but the polypeptide chain overlaid well with other models (Table 4-9). Overall, the quality of the *Native* X-ray structure is the poorest among others. It can be seen from the Figure 4-19(b) that β3 strand was not identified correctly using either DSSP (Kabsch and Sander, 1983) or PROCHECK (Laskowski *et al.*, 1993).

As will be described in Chapter 5, water molecules are pivotal to mediate contact of the MeCP2 MBD domain and the methyl groups. The best refined structural model with the highest number of water molecules was selected for detailed structural analysis: *A140SeMet-Mn* has the highest resolution data, the best refined structure and most importantly, contains the highest number of water molecules (total 47 water molecules). The quality of model *Iodo17* is generally good but only 30 water molecules were assigned due to a relatively low resolution data. In order to ensure that the *A140SeMet-Mn* is identical in particular to the native X-ray structure, the models were superimposed using either C_α or all atoms (Table 4-9). Overlay of *A140SeMet-Mn* and *Native* using all C_α produced an RMSD fit of 0.43Å. This indicates that both structures are essentially identical (Figure 4-20). In fact, all models are identical to *A140SeMet-Mn* with C_α RMSD fit below 0.6Å. Nevertheless,

overlaying using all atoms except water molecules give RMSD fit generally higher than the C_α alone.

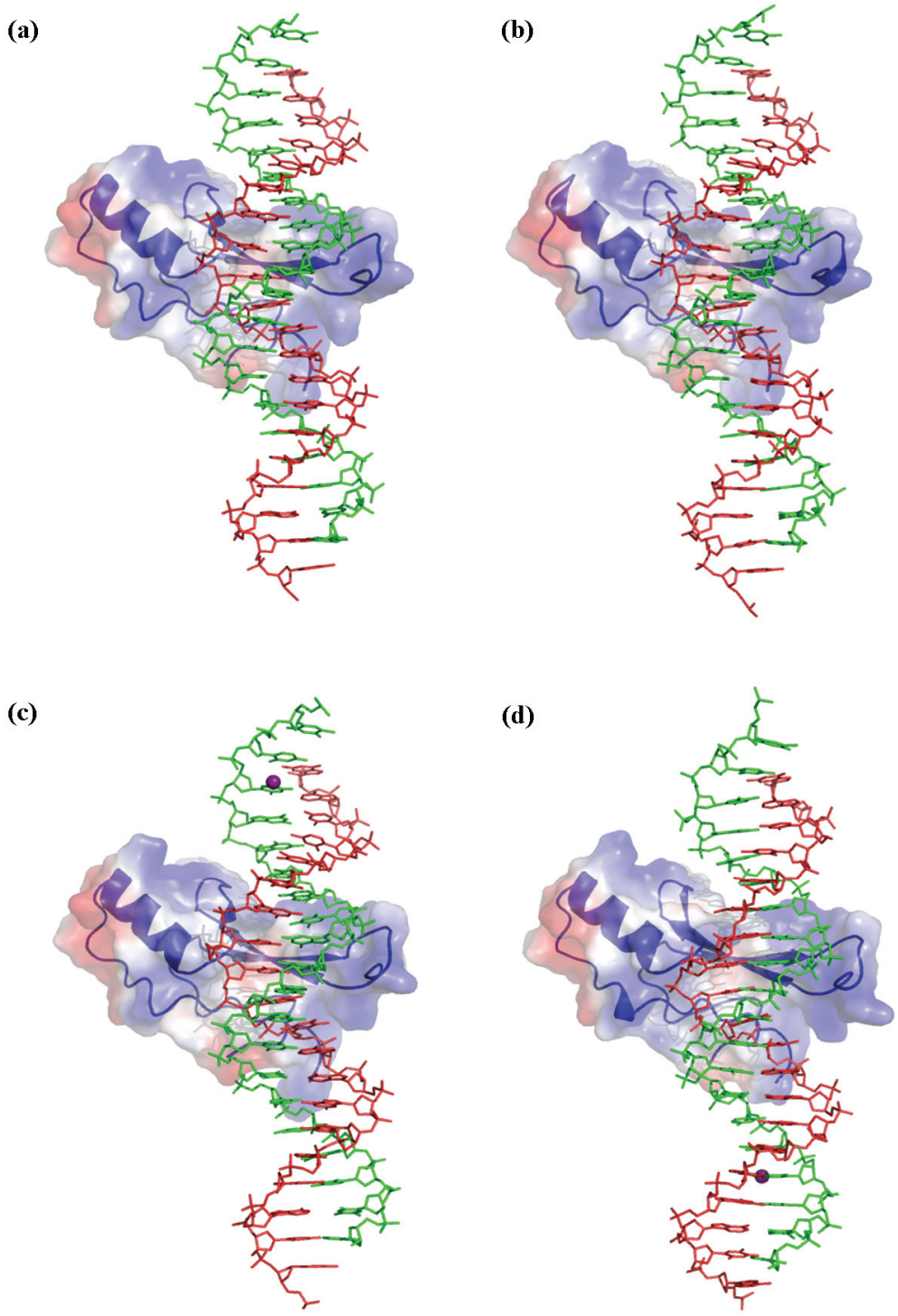


Figure 4-19 Refined models of *A140SeMet-Mn*, *Native*, *Iodo3* and *Iodo17*
(a) X-ray structure of *A140SeMet-Mn* from selenium *Peak* data, (b) *Native* X-ray structure, (c) iodinated derivative *Iodo3* and, (d) iodinated derivative *Iodo17*.

Table 4-8 Refinement statistics

Crystal	***A140SeMet***	***Native***	***D21***	***D22***
Dataset	*A140SeMet-Mn*	*Native*	*Iodo17*	*Iodo3*
Refinement				
R_{cryst}/R_{free} (%)	21.2/ 27.6	23.0/ 29.5	24.3/30.1	24.9/29.5
Geometry				
rmsd bond/angles (Å, °)	0.009/1.85	0.016/2.67	0.010/1.901	0.017/2.917

R_{free} is R_{cryst} calculated for a test set of randomly chosen 5% of the data.

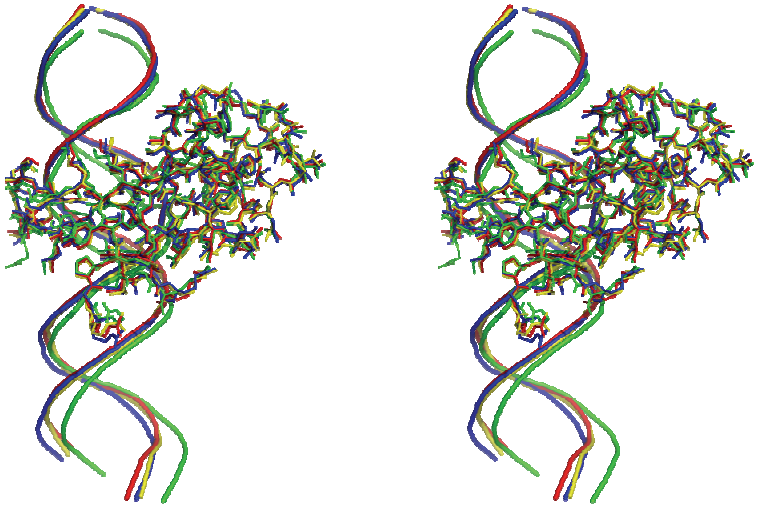


Figure 4-20 Overlay A140SeMet-Mn with native and iodinated X-ray structures
Stereo views of overlay of *Iodo17* (green), *Iodo3* (blue) and *Native* (yellow) models onto *A140SeMet-Mn* using all C_α atoms. All models show identical structure as the best refined *A140SeMet-Mn* except iodinated DNA duplex from *Iodo17*.

Table 4-9 RMSD fit of all 4 structures using C_α and all atoms

	Native	***Iodo17***	***Iodo3***
A140SeMet-Mn	0.43 (0.62)	0.59 (0.98)	0.44 (0.67)
Native	-	0.77 (1.13)	0.49 (0.76)
Iodo17	-	-	0.70 (1.09)

Numbers in parenthesis are RMSD values calculated using all atoms

CHAPTER 5. STRUCTURAL ANALYSIS

5.1 INTRODUCTION

The solution structure of MBD domains from MeCP2 and MBD1 were published in 1999 (Ohki *et al.*, 1999; Wakefield *et al.*, 1999). These NMR structures generally explored the tertiary protein folding of the MBD domain and the potential residues that are involved in the DNA binding. Two years later, Ohki *et al.* reported the first structure of MBD protein in complex with a methylated DNA (Ohki *et al.*, 2001). The authors claimed that the binding specificity of the MBD domain depends on hydrophobic interactions between the methyl groups of m5C and a hydrophobic patch comprising of 5 residues, Val20, Arg22, Tyr34, Arg44 and Ser45, within the MBD domain (Ohki *et al.*, 2001). The X-ray structure determined in this study reveals, however, that the methyl groups make contacts with a predominantly hydrophilic surface that includes 5 water molecules.

The Rett missense mutations occur throughout the *MECP2* gene with more than half of disease-causing Rett missense mutations affecting the core MBD between amino acid 97 and 161, although this constitutes only approximately 13% of the total amino acid sequence. Of these; T158M, R133C, R106W account for 73% of Rett mutations found within the MBD domain (Kriaucionis and Bird, 2003). The crystal structure highlights the role of these 3 important residues in stabilising protein folding and in DNA-protein interactions. The X-ray structure also explains how T158 (the most common missense mutations of RTT) coordinates the unusual Asx-ST motif. This motif involves two consecutive turns at the C-terminal region of the MeCP2 MBD domain.

Recently discovered AT bases close to the symmetrical methyl-CpG dinucleotides that promotes high binding of MeCP2 (Klose *et al.*, 2005) could possibly contribute to DNA bending. The structure presented here shows hydrogen bonds connecting Val159(N) to the phosphate group of T31. The AT bases also display a high degree of propeller twist and distinct features of narrowing the minor groove and widening of major groove. These features may have a consequence of DNA bending and stability. However, the roles of the AT run in increasing MeCP2 binding remain unknown.

The X-ray structure from the crystal *A140M-SeMet-Mn* (with the lowest R_{cryst}/R_{free}, RMSD bond/angle and highest FOM) (Table 4-8) was used for full structural analysis using various programmes. This structure rationalises the effects of the common Rett mutations and provides a new and general model for methylated-DNA binding that is dependent on structured water molecules discovered in this study.

5.2 MATERIALS AND METHODS

The final model from the selenium peak data was used for structural analysis. Multiple sequence alignment was performed with T-coffee (O'Sullivan *et al.*, 2004). The secondary structure and geometry features were defined using DSSP (Kabsch and Sander, 1983) and PROMOTIF (Hutchinson and Thornton, 1996) and displayed with ESPript (Gouet *et al.*, 1999). Structural motifs of the protein were analysed with MSDMotif (Golovin *et al.*, 2004) and the potential hydrogen bonds were predicted with HBPLUS (McDonald and Thornton, 1994). The X-ray DNA geometry was analysed with 3DNA (Lu and Olson, 2003) and all figures with structural representatives (eg. lines, sticks, spheres, cartoons and electron density) were prepared using Pymol (DeLano, 2003).

5.3 RESULTS AND DISCUSSION

5.3.1 Overall structure

To obtain a high resolution crystal structure of the MeCP2 MBD-DNA complex, the recent discovery of 4 or more AT bases close to the central methyl-CpG which promotes high affinity binding of MeCP2 (Klose *et al.*, 2005) was taken into consideration. The recombinant protein of the MeCP2 MBD domain (construct 77-167) contains 91 amino acids (Figure 5-1), immediately followed by an uncleavable 6xHis tag at the C-terminus. This was co-crystallised with the synthetic nucleic acid comprising a 20 bp DNA fragment (nucleotides -108 to -90) of promoter III of the *BDNF* gene, which contains a central methyl-CpG pair and an adjacent four AT run (Figure 5-2a). This promoter was implicated as an MeCP2 endogenous target gene (Chen *et al.*, 2003; Martinowich *et al.*, 2003).

Structural analysis of the best model (*A140SeMet-Mn*) reveals that 75 % of the amino acids can be defined from the electron density map. Unfortunately, the first 14 residues at the N-terminal region (amino acid 77-90) and the last 11 residues at the C-terminal region (amino acid 163-167 and the 6xHis tag) however could not be traced at this resolution (2.5 Å). This was largely due to the floppy loops located at both C- and N-terminals of the construct 77-167. All 40 nucleotides can be defined as a slightly bent B-form Watson-Crick double-stranded DNA helix. In addition, a total of 47 water molecules were also identified. Although Mg^{2+} or Mn^{2+} was incorporated into the crystallisation solution, none of these ions were identified in the X-ray structure. The overall structure of the MBD domain in complex with methylated DNA is shown in Figure 5-2b and c. The protein and DNA are represented by cartoon (chain A) and sticks (chains B and C), respectively, and the protein-DNA interactions mainly occur within the major groove which contains the m5C methyl groups. The single-base overhangs of 5'T1 and 5'A21 on chain B and C, respectively, form end-to-end base stacking with the adjacent double helices. The actual sequence corresponding to the nucleotides -108 to -90 of promoter III of mouse *BDNF* gene is the complementary 19 base pair DNA region.

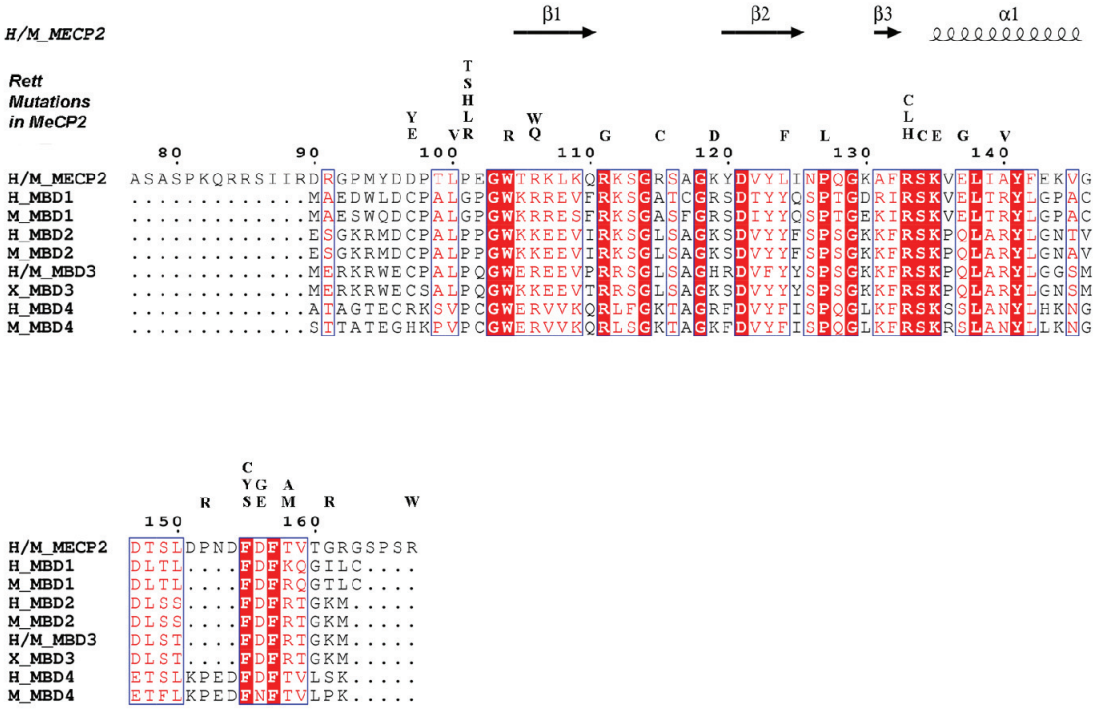


Figure 5-1: Sequence comparison of methyl binding domain of the MBD protein family from Human (H), mouse (M), and *Xenopus* (X) MeCP2 MBD
Secondary structure elements from the X-ray structure are denoted by arrows (β-strand) and coils (α-helix). Mutations that have been associated with neurological disease are shown.

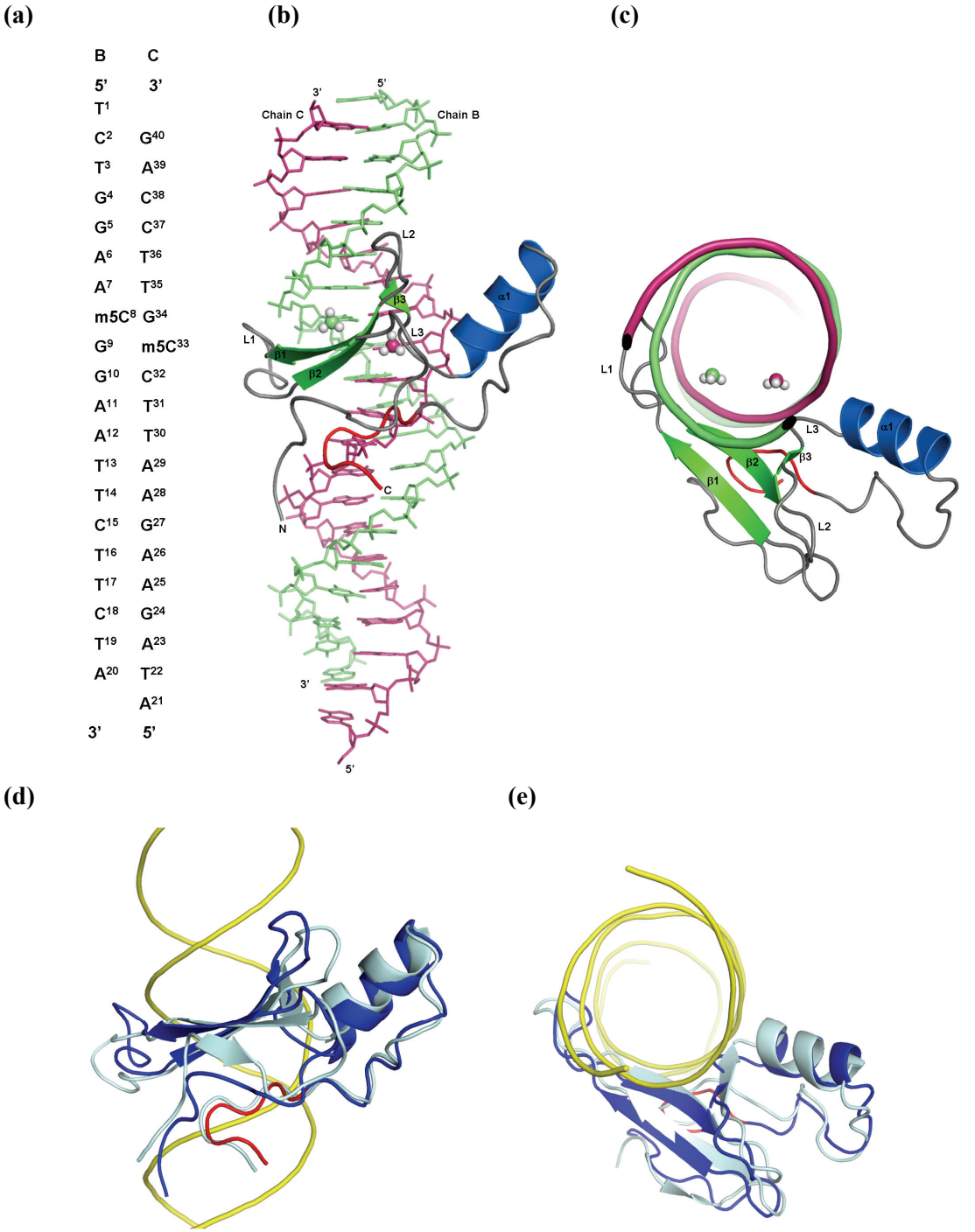


Figure 5-2: The conformation of the X-ray structure of MeCP2-MBD complexed with *BDNF* promoter DNA at 2.5 Å is similar to the unliganded MBD

(a) *BDNF* promoter DNA sequence used in the co-crystal X-ray structure. The overhanging T/A bases were incorporated to promote crystallisation through end-to-end DNA stacking. (b) The X-ray structure of MeCP2-MBD in complex with DNA. The methyl groups of the mCpG pair are shown as spheres. The β-strands (green) and α-helix (blue) are connected by the loops L1 (R111- K119), L2 (N126- K130) and L3 (R133- V135). The tandem Asx-ST motif (D156- R162) is highlighted in red. (c) X-ray structure viewed along the DNA helix. (d) Overlay of the unliganded MeCP2 structure determined by NMR (grey) with the X-ray structure (blue) in complex with DNA (yellow). The Asx-ST turn of the X-ray structure is highlighted in red and (e) a view of the overlay along DNA double helix as (c).

5.3.2 Secondary structure and protein folding

The secondary structure of the MeCP2 MBD domain in complex with 20 bp methylated DNA was examined using the programmes DSSP (Kabsch and Sander, 1983) and PROMOTIF (Hutchinson and Thornton, 1996). Both programmes defined a consistent secondary structure as shown in Figure 5-1. In general, the MBD domain contains three β-sheets which consist of 19.4% while the α-helix accounts for only 15.3% of total amino acids (Table 5-1). The secondary structure determined in this X-ray study is similar to that of the solution structure of MeCP2 MBD alone (Wakefield *et al.*, 1999) except an additional β-sheet (amino acid ^{96}DDP^{98}) at the N-terminus defined by Wakefield and coworkers (1999) (Figure 5-2c). Overlaying the X-ray and NMR determined MBD domains shows that both structures are similar with an RMSD fit for C_α atoms of 2.33 Å (Figure 5-2d, e). The most notable difference concerns loop L1, which is drawn towards the DNA through four hydrogen bonds with the phosphate backbone on the surface of the major groove that includes the pair of cytosine methyl groups (Figure 5-2d). A subtle conformational change also occurs in Loop L2 where the loop is brought closer to the DNA although no hydrogen bond are observed.

Table 5-1 Secondary structure of X-ray MeCP2 MBD
Structure analysed with PROMOTIF (Hutchinson and Thornton, 1996)

Secondary structure	Range (number of residues)	Amino acid sequence
N-terminal loops	91-104 (14)	RGP(Mse)[†]YDDPTLPEGW
β1	105-110 (6)	TRKLKQ
L1	111-119 (9)	RKSGRSAGK
β2	120-125 (6)	YDVYLI
L2	126-130 (5)	NPQGK
β3	131-132 (2)	AF
L3	133-134 (2)	RS
α1	135-145 (11)	KVELI(Mse)[‡]YFEKV (16.70 Å)*
C-terminal loops	146-162 (17)	GDTSLDPNDFDFTVTGR

[†] in the wild type sequence, this position consists of Met94
[‡] in the wild type sequence, this position consists of Ala140
*length of the α-helix in Angstrom

The crystal structure of the MeCP2 MBD complexed with DNA also displays a high structural homology with the secondary structure of NMR established MBD1 MBD-methylated DNA complex published by Ohki and coworkers (Ohki *et al.*, 2001). Overlay using the 20 C_α atoms from strand β1, β2 and α1 gives an RMSD fit of 1.4 Å (Figure 5-3a). This superposition shows that, both DNA duplexes are not in a similar orientation (Figure 5-3b) despite the similarity of their binding proteins. Though the MBD1 is folded in a similar way, there are some important differences in the molecular details of recognition and binding. Using ensemble 1 of the MBD1 MBD-DNA complex (pdb: 1IG4) for PROMOTIF analysis, the length of the α-helix comprising of 8 residues is 4.7 Å shorter than MeCP2 MBD α-helix. In the C-terminal region, MBD1 has a 4-residues (β-turn ^{151}DPND^{154} in MeCP2) deletion and has a type I β-turn for amino acids ^{63}DFKQ^{66} (Figure 5-1) which cannot adopt an equivalent Asx-ST-motif as in the MeCP2 MBD domain. MeCP2 MBD is hydrogen bonded to the AT run DNA phosphates which may play a role in complex formation. It is possible that the absence of this unique structural motif or an AT run close to methyl-CpG resulted a different DNA orientation for the NMR established MBD1 MBD-DNA complex.

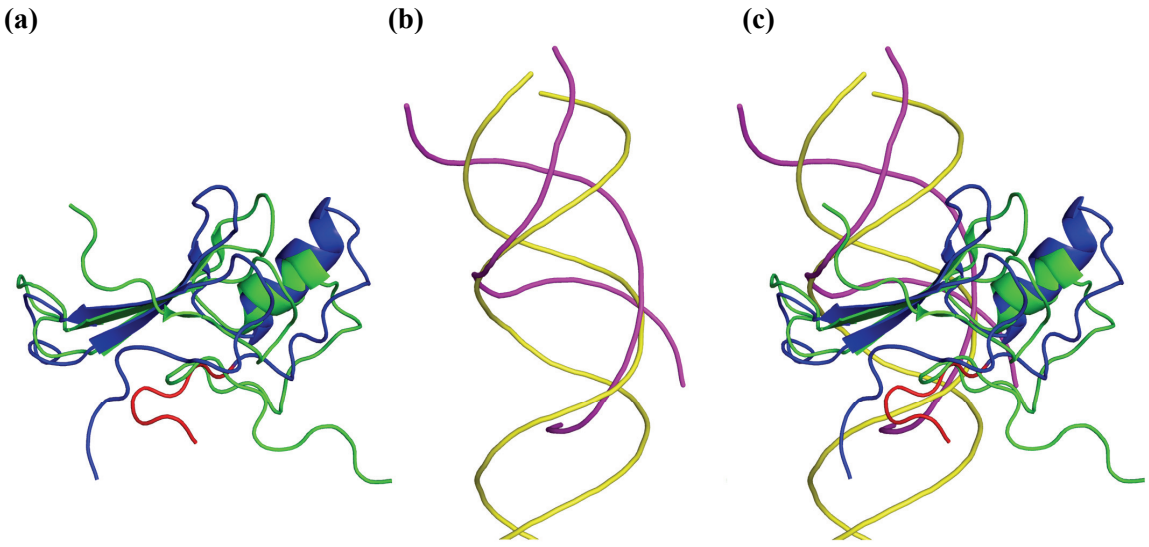


Figure 5-3 Overlay of MeCP2 (X-ray) and MBD1 (NMR) MBD-DNA complexes
(a) superposition of the MBD domains of MeCP2 (blue and red) and MBD1 (green) using 20 C_α atoms from strands β1, β2 and α1 gives an RMSD fit of 1.4Å. Asx-ST-Turn is coloured in red. (b) DNA positions resulted from overlaying of their binding proteins. The *BDNF* and 12 bp DNA are represented in yellow and magenta, respectively. (c) overlay of the complexes of MBD1 and MeCP2 MBD with their methylated DNA.

Analysis with PROMOTIF also highlights other characteristics of the MBD domain of MeCP2. All together 9 β-turns have been identified (Table 5-2). The first 2 consecutive β-turns (^{97}DPTL^{100} and ^{101}PEGW^{104}) preceding the first β-strand of the protein structure. Interestingly, Loop L1 contains 3 β-turns: ^{113}SGRS^{116} (type IV), ^{114}GRSA^{117} (type I) and ^{116}SAGK^{119} (type II). L2 composed of type I β-turn (^{126}NPQG^{129}). As shown in Figure 5-1, the β-turns composed of ^{151}DPND^{154} before the conserved FDF motif is only present in MeCP2 and MBD4 but not other MBD proteins. This indicates that the DNA binding mode of MBD4 might closely resemble MeCP2. However, the requirement of AT bases for high affinity binding of MBD4 is still unexplained. This type I β turn could possibly position the unique structural motif proximal to the DNA phosphate backbone.

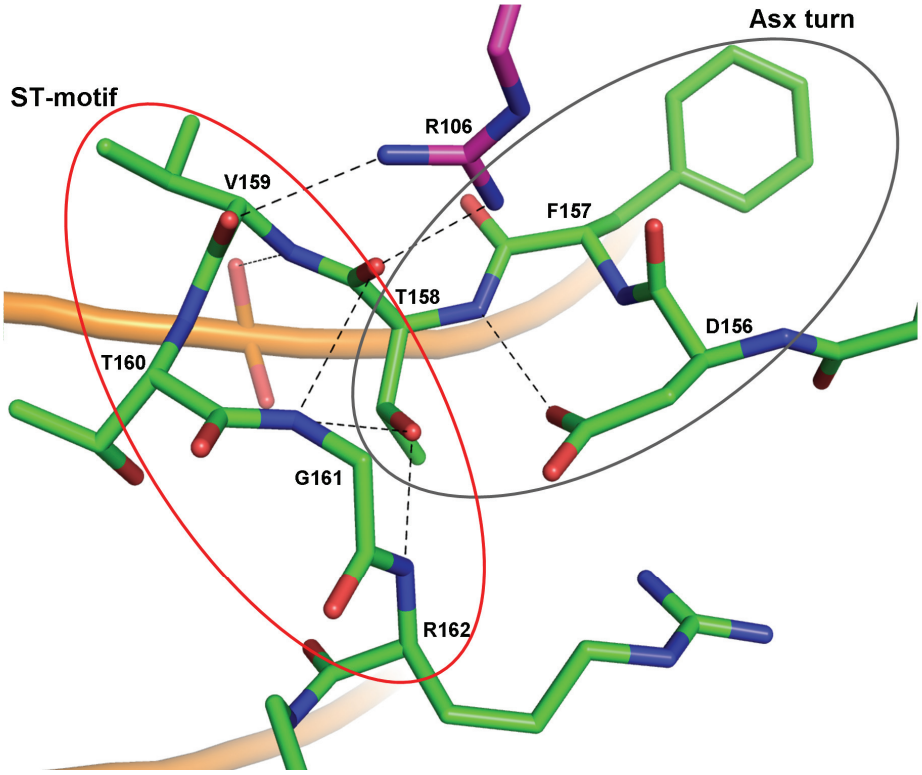
Table 5-2 MeCP2 MBD β-turns
Structure analysed with PROMOTIF (Hutchinson and Thornton, 1996)

Residues	Sequence	Type	*i* to *i*+3 distance (Å)
97-100	DPTL	I	5.8
101-104	PEGW	II	5.7
113-116	SGRS	IV	6.6
114-117	GRSA	I	5.4
116-119	SAGK	II	5.6
126-129	NPQG	I	5.0
151-154	DPND	I	5.7
152-155	PNDF	IV	5.9
158-161	TVTG	I	5.1

5.3.3 Unique roles of T158 in tandem Asx-ST-motif

Mutation of T158M, R133C and R106W account for the highest number of Rett cases (Kriaucionis and Bird, 2003). Using the MSDmotif server (Golovin *et al.*, 2004), the structural roles of Thr158 and Arg106 in stabilising a unique structural motif; which comprises of two consecutive turns; have been defined (Figure 5-4a). The first structural motif is an Asx motif and comprises ^{156}DFT^{158} followed immediately by an ST-β-turn (^{158}TVTG^{161}). This structural motif was named as a 'tandem Asx-ST-motif' (Ho *et al.*, 2008).

(a)



(b)

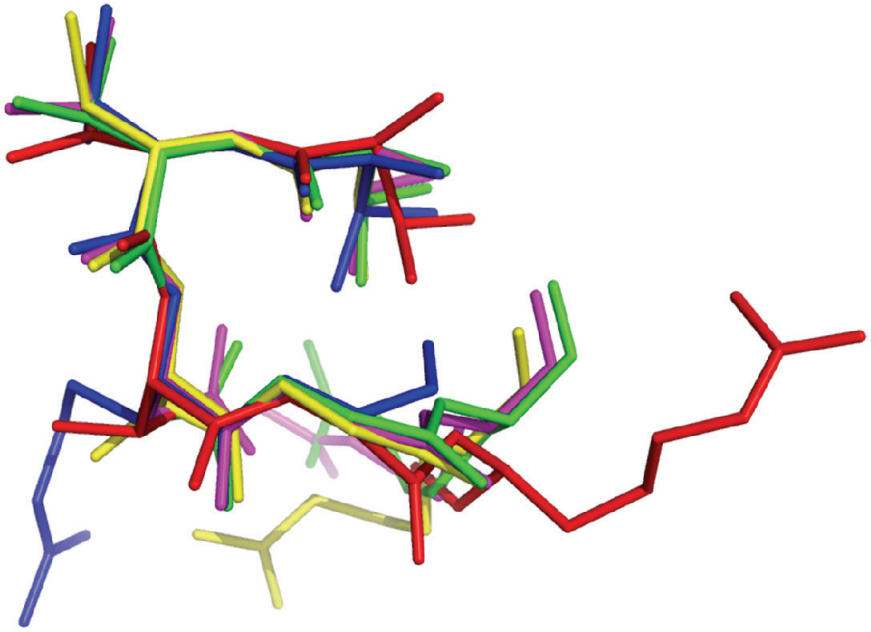


Figure 5-4: T158 plays a structurally important role in forming the tandem Asx-ST motif.

(a) The Asx turn is composed of D156, F157 and T158 with the motif-defining hydrogen bond formed between the D156 carboxylate side-chain and the main-chain amine nitrogen of T158. The ST-motif is the modified β-turn consisting T158, V159, T160 G161 and R162 with the side chain hydroxyl group of T158 hydrogen bonding to the main chain nitrogen atoms of G161 and R162. (b) Overlay of similar structural motifs identified from Protein Database (Berman *et al.*, 2000), TVDGR (1CBS, green), TVRG (1PTA, blue), TVDGR (1XCA, magenta), and TVTGR (4KIV, yellow), with the X-ray structure of this study (red).

By definition, the Asx-motif is characterised by three amino acids with residues *i* being either Asp or Asn with the carboxylate side-chain forming a hydrogen bond with the main-chain amine group of the third (*i+2*) amino acid (Wan and Milner-White, 1999a). The ST-β-turn is characterised by a β-turn with Ser and Thr in position of residue *i* with the side-chain hydroxyl further stabilising the turn with a hydrogen bond to NH of *i*+3 and *i*+4 residues (Wan and Milner-White, 1999b). The Asx turn is formed by a hydrogen bond that connects the main-chain NH of Thr158 and the side-chain carbonyl of Asp156 (Table 5-3 and Figure 5-4). The ST-turn is shaped by two hydrogen bonds connecting the side-chain hydroxyl group of Thr158 to the main-chain amide groups of Gly161 and Arg162. Thus, Thr158 occupies a pivotal position that coordinates the two consecutive turns. Interesting, this tandem Asx-ST-motif is further stabilized by hydrogen bonds to Arg106 which connects the main-chain carbonyl group of Thr158 and Val159 (Figure 5-4a).

Table 5-3 Hydrogen bonds in Asx-ST-motif of MeCP2 MBD domain

Motif	Residue	Atom	Residue	Atom	Distance (Å)
Asx turn	Asp156 (*i*)	OD2	Thr158 (*i+2*)	N	2.9
ST-β-turn	Thr158 (*i*)	OG1	Gly161 (*i+3*)	N	3.1
		OG1	Arg162	N	2.7
		O	Gly161 (*i+3*)	N	3.3
		O	Arg106	NH2	2.6
	Val159 (*i+2*)	N	T31	O2P	3.2
		O	Arg106	NH1	2.9

In agreement with the PROMOTIF (Hutchinson and Thornton, 1996) analysis, MSDmotif analysis shows that ^{158}TVTG^{161} belong to type I *β*-turn. The *φ*/*ψ* angles of *i*+1 (Val159) and *i*+2 (Thr160) are -55.1/-57.0˚ and -71.3/-19.4˚, respectively (Table 5-4), and this fulfilled the requirements for a type I β-turn. Small structural motif searches using the MSDmotif server reveals that there are 1773 X-ray structures in the Protein Database (Berman *et al.*, 2000) with resolution better than 2.5Å containing an equivalent β-turn, of which, only 4 bear the sequence TVXG (Table 5-4). Overlay of these structures with the ST-β-turn in this study gave C_α RMSD fits between 0.24 and 0.49Å (Table 5-4 and Figure 5-4b). This highlights the defined conformation of the type I β-turn composed of TVXG as observed in this study.

Table 5-4 Overlaying of type I β-turns with ^{158}TVTGR^{162} of MBD in this study

pdb	References	Motif	Residue *i*+1 φ/ψ*	Residue *i*+2 φ/ψ*	RMSD fit/Å
This study		TVTGR	-55.1/-57.0	-71.3/-19.4	-
1CBS	(Kleywegt *et al.*, 1994)	TVDGR	-56.88/-37.13	-83.47/6.03	0.31
1PTA	(Benning *et al.*, 2001)	TVRG	-69.22/-10.36	-115.67/5.48	0.24
1XCA	(Chen *et al.*, 1998)	TVDGR	-50.79/-19.01	-102.23/0.83	0.40
4KIV	(Mochalkin *et al.*, 1999)	TVTGR	-51.68/-19.71	-99.26/17.35	0.49

* To be classified as a type I β-turn, the φ and ψ angles of the residues *i*+1 and *i*+2 must be: $-140 < \varphi < -20$ and $-90 < \psi < +40$. The Cα RMSD fits were calculated by overlaying the selected motifs on TVTGR of this study.

The location of this Asx-ST motif is not influenced by the bound DNA as it is invariant from the unliganded form (Wakefield *et al.*, 1999). The conformation of the Asx and ST turn seems crucial in the interaction of MeCP2 with the DNA. The amine group of Val159 forms a hydrogen bond with atom O2P of T31 (see Table 5-3 and Figure 5-4a), which coincides with the start of the AT-run (nucleotide T31 on chain C). It is known from binding studies that MeCP2 (in contrast with MBD1) requires a proximal AT run for high affinity binding (Klose *et al.*, 2005). The MBD1 sequence is shorter by four residues (β-turn ^{151}DPND^{154} in MeCP2) in this region and cannot adopt an equivalent Asx-ST-motif shape. Therefore, it is hypothesised that the interaction of the Asx-ST-motif is important in the recognition of the proximal AT run by MeCP2. The missense mutation of T158M and R106W may destabilise this tandem Asx-ST-motif and subsequently disrupt the DNA binding. Mutational analysis of T158 will be discussed in Chapter 6.

5.3.4 Interactions of MeCP2 MBD domain and *BDNF* fragment

The DNA-protein interactions occur exclusively within the methyl-m5C containing DNA major groove, involving three base pairs (Figure 5-5a), one above and one below of methyl-m5C, at both strands. The MBD domain makes contacts with the DNA by direct and water molecules mediate protein side-chain interactions with the DNA bases and phosphate backbone. Surprisingly, only a minimum direct contact between the protein and the symmetrical methy-m5C dinucleotides was observed. The DNA-protein interaction analysis reveals that contacts between the MBD and methyl groups of m5C indeed depend upon a specific hydration pattern at the major groove of

the DNA. In contrast to the unmodified cytosine, the methylated cytosine bases are more hydrated.

5.3.4.1 Water mediated interactions

The contacts between the methyl-CpG binding surface of MBD1 MBD and the C5 methyl groups were thought to be hydrophobic interactions involving a hydrophobic patch comprises of Val20, Arg22, Tyr34, Arg44 and Ser45 (Figure 5-1) (Ohki *et al.*, 2001). In contrast to Ohki *et al.* (2001), the methyl-CpG-binding surface in this study is unexpectedly hydrophilic. The C5 methyl groups (m5C8 and m5C33) form close contacts (less than 4Å) with 13 and 12 neighbouring atoms respectively (Figure 5-5a and Table 5-5). Of the 25 interactions only two (m5C33 to CG and CB of Arg133) are classically hydrophobic in character. MeCP2 recognition of methyl-CpG involves the five water molecules W12, W24, W10, W21 and W22, each making a C-H…O interaction with one of the methyl groups (Figure 5-5a and b). W24 and W22 form a tetrahedral arrangement of hydrogen bonds that bridge DNA with protein and also C-H…O interactions with both methyl groups of the methyl-CpG dinucleotide pair (Figure 5-5a and b). W22 forms four hydrogen bonds with, Asp121, W24, W21 and N4 of m5C33 plus a C-H…O interaction with the C5 methyl group of m5C33. W24 forms four conventional hydrogen bonds with Tyr123, Arg133, W22 and N4 of m5C8 plus a C-H…O interaction with the methyl group of m5C8 (Figure 5-5a, b and Table 5-5). Interestingly, C-H…O hydrogen bonds have been implicated in methyl group recognition during lysine methylation by SET domain lysine methyltransferases (Couture *et al.*, 2006). The importance of Asp121 in maintaining the DNA-protein interaction has been tested using site-directed mutagenesis and the result is presented in Chapter 6.

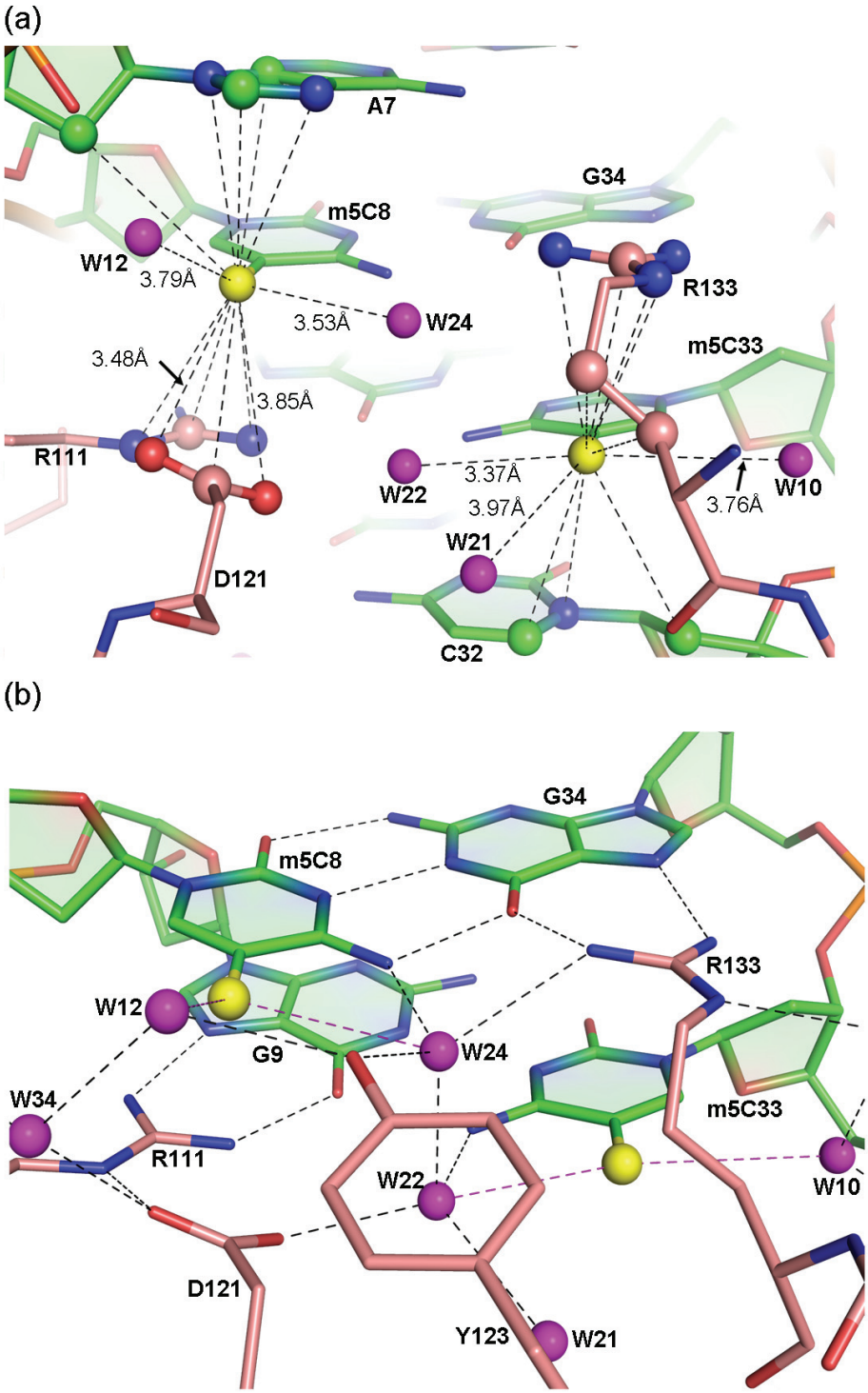


Figure 5-5 m5C methyl groups interactions

(a) The m5C methyl groups are shown as yellow balls. Non-bonded contacts to the m5C methyl groups of less than 4Å are shown as black dashed lines. The water molecules W10, W12, W21, W22 and W24 are drawn as purple balls.

(b) Hydrogen bonds are shown as black dashed lines. The two m5C methyl…water contacts are drawn as dashed red lines. Distances for all hydrogen-bonds shown in this figure are tabulated in Table 5-5. Specific hydrogen bonds formed between R111 and G9 (chain B) and R133 and G34 (chain C) position the guanidinium groups directly over the methyl groups of the m5C bases.

Table 5-5 Hydrogen bonds and van der Waals contacts at the methyl-m5C recognition surface.

Residue/ nucleotide	Atom	Residue/ nucleotide	Atom	Distances (Å)
m5C8	methyl	A8	C2*	3.9
			N9	3.7
			C8	3.6
			N7	3.8
			C4	3.9
		Arg111	NE	3.9
			CZ	3.9
			NH2	3.6
		Asp121	OD1	3.9
			OD2	3.5
			CG	3.8
		W24		3.5
		W12		3.8
m5C33	methyl	C32	C2*	3.9
			N1	3.9
			C6	3.8
		Arg133	NH1	3.9
			NH2	3.7
			CZ	3.5
			NE	3.7
			CG	3.7
			CB	3.4
		W21		3.9
		W22		3.4
		W10		3.8

5.3.4.2 Interactions of MBD-DNA bases

The only MeCP2-MBD residues that directly interact with DNA bases in the crystallised *BDNF* sequence are Asp121, Arg111 and Arg133 (Figure 5-5a, b and Table 5-5). Asp121 makes a CH…O hydrogen bond of 3.5Å with the methyl group of m5C8 (Figure 5-5a). The hydrogen bonds formed between the symmetrical arginine fingers (Arg111 and Arg133) and each guanine of the methyl-CpG pair (Figure 5-6) occur frequently in diverse examples of protein-DNA recognition (Luscombe *et al.*, 2001). Both arginine fingers lie in a plane with the guanine bases and are locked in position by salt bridges with the carboxylates of Asp121 and Glu137 (Table 5-6).

Table 5-6 Water mediated interactions at MeCP2 MBD-DNA contact interface

Residue/ nucleotide	Atom	Residue/ nucleotide	Atom	Distances (Å)
W24		m5C8	N4	2.7
			methyl	3.5
		Arg133	NH2	3.1
		Tyr123	OH	2.9
		W22		2.6
W22		m5C33	N4	2.9
			methyl	3.4
		Asp121	OD1	2.7
		W21		3.0
		W24		2.6
W12		m5C8	methyl	3.8
		W34		2.8
		Tyr123	OH	2.7
		A7	O2P	2.7
W10		m5C33	methyl	3.8
			O2P	2.6
		W42		3.3
Tyr123	OH	W12		2.7
		W24		2.9
Asp121	OD1	m5C8	methyl	3.9
		W22		2.7
		Arg111	NH2	2.8
	OD2	m5C8	methyl	3.5
		W34		3.0
		Arg111	NE	2.7
	CG	m5C8	methyl	3.8
Arg111	NE	m5C8	methyl	3.9
		Asp121	OD2	2.7
	NH2	m5C8	methyl	3.6
		Asp121	OD1	2.8
	CZ	m5C8	methyl	3.9
Arg133	NH1	m5C33	methyl	3.9
	NH2	m5C33	methyl	3.7
		W24		3.1
	NE	m5C33	methyl	3.7
		Glu137	OE1	2.8
	CB	m5C33	methyl	3.4
	CG			3.7
	CZ			3.5

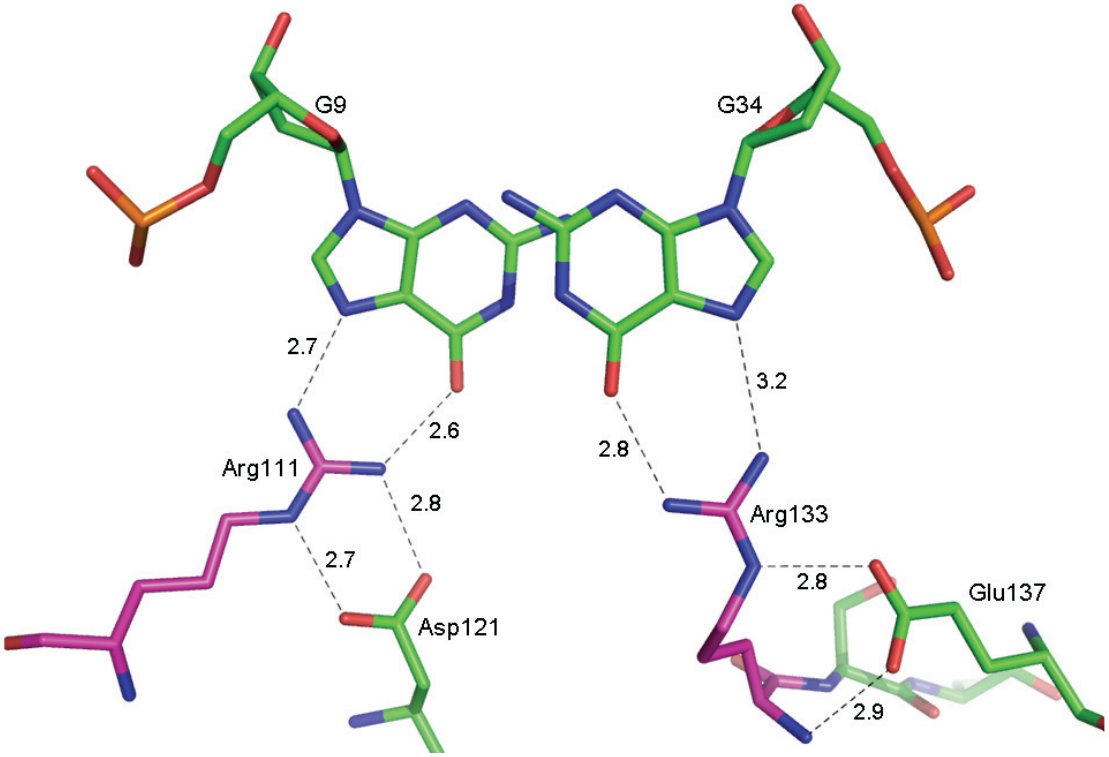


Figure 5-6 Arginine fingers
Both Arg111 and Arg133 are held in place by salt-bridge interactions with from Asp121 and Glu137, respectively. All the classical hydrogen bonds are within 3.5 Å.

This symmetrical arrangement of the arginine side chains places the guanidinum groups directly above/beneath the methyl groups of the methylated cytidine bases with an average methyl-guanidinium distance of 3.7Å (Figure 5-5a). Specificity for the methyl-CpG base pair derives in part from the constrained configuration of arginine side-chains in contact with the juxtaposed guanines that are exclusive to the CpG sequence motif. These direct base contact residues are conserved in the MBD family (Figure 5-1), therefore, it is speculated other MBD proteins also recognise their endogenous target DNA with these residues.

One of the key focuses of this study is to determine the methyl-CpG binding specificity of the MeCP2 MBD domain. The presence of structurally conserved water molecules due to cytosine methylation in the major groove possibly establishes the methylated DNA binding specificity of the protein. This methylation-specific hydration pattern has been noted previously and was speculated to potentially specify recognition of methylated DNA by proteins (Mayer-Jung *et al.*, 1998). Mayer-Jung and co-workers reported that the m5C modified the hydration pattern surrounding the methyl groups by attracting more water molecules. Two water molecules connect the

methyl group of m5C through CH…O hydrogen bonds, and these water molecules also individually hydrogen bond to N4 and the m5C phosphate oxygen (Mayer-Jung *et al.*, 1998). In the X-ray structure of this study, two equivalent water molecules were also observed for each m5C. Both W24 and W12 are C-H…O bonded to methyl group of m5C-8 but separately connect the O2P of A7 and N4 of m5C-8 (Figure 5-7a). In this MeCP2 MBD-DNA complex, W22 and W10 behave similarly by connecting the methyl group of m5C-33 through C-H…O hydrogen bonds and individually bond with N4 and O2P of the same nucleotide (Figure 5-7b). In addition to other water molecules in the major groove, it is speculated that the string of waters, composed of W34, W12, W24, W22, W21, and W10, at the DNA-protein interface (Figure 5-5) results from cytosine methylation. In contrast, unmodified cytosine bases in this X-ray structure are less hydrated than the methylated cytosine and the hydration pattern due to the C5 methylation-specific characteristic has not been shown.

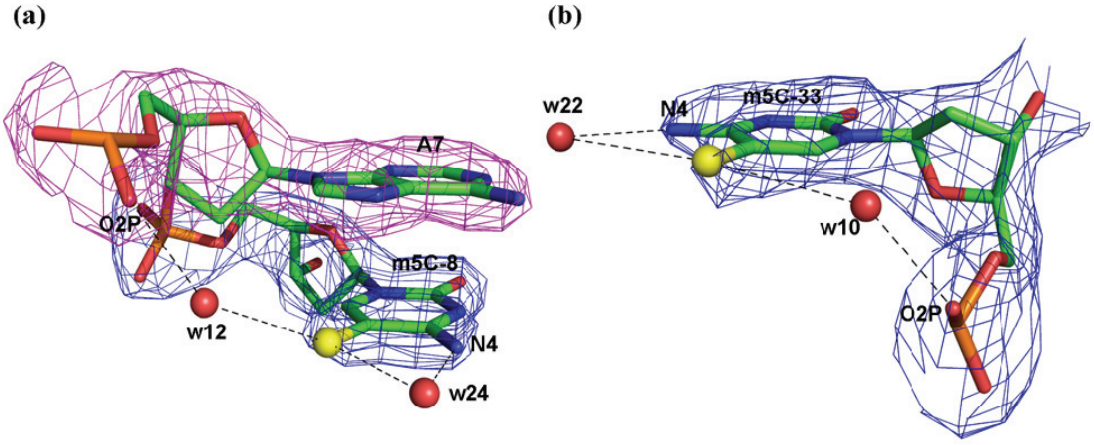


Figure 5-7 Methylation specific hydration of methyl-CpG
(a) Both W24 and W12 interacting with methyl group of m5C-8 though C-H…O hydrogen bonds but separately contacting N4 of the same nucleotide and O2P of A9.
(b) Both W22 and W10 forming C-H…O hydrogen bonds with methyl group of m5C-33. W10 also mediating the contact of methyl and O2P of the same nucleotide. The electron density for m5C and A9 are coloured in blue and pink, respectively. Methyl groups are represented by yellow spheres.

This observation was further supported by analysing the deposited X-ray determined DNA structures in the Nucleic Acid Database (NDB) (Berman *et al.*, 1992). The NDB contains all types of nucleic acid structures including ligands and protein bound complexes determined using X-ray crystallography or NMR (Berman *et al.*, 1992). All m5C containing DNA structures were mined using the NDB integrated searches.

Various searches interrogating the effect of nucleic acid modifications, nucleic acid conformational type and components of biomolecules were carried out (Berman *et al.*, 1992). In the initial round, by ignoring the base and sugar modifications, all together 3658 hits were found. By choosing double-stranded DNA structures only and solved by X-ray crystallography, regardless of their conformational types, the positive candidates were reduced to 656 hits. Further elimination using 3Å resolution cut off yielding 649 hits. Out of these structures, 47 contained m5C; all water molecules within 4Å from the methyl groups were identified. The water molecules that bridge the DNA phosphate group to the methyl group and those forming CH...O bonds with the methyl group and simultaneously connect the N4 of 5mC were identified. Of these, 78% have at least one m5C base coordinating a water molecule through a CH…O hydrogen bond to the methyl carbon and simultaneously a conventional hydrogen bond to N4 (similar to the interactions shown by W22 and W24) and with this characteristic, 84% contain at least one water molecule bridging the DNA phosphate group to the methyl group of m5C (Figure 5-7).

5.3.4.3 Interactions of MBD and DNA phosphate backbone

The major MBD-DNA base contacts involve the recognition of symmetrical methyl-CpG dinucleotides in the major groove of the crystallised *BDNF* fragment. Apart from the MBD-DNA base contacts, there are 7 additional classical hydrogen bonds; mainly composed of Ser and Lys residues; connecting the MeCP2 MBD domain to the DNA phosphate backbone. Figure 5-8 shows that 4 out of 7 DNA backbone contacts are located on loop L1. These hydrogen bonds draw the loop L1 to the DNA phosphate groups. The main-chain amine group of Lys112 and side-chain of Ser113 are hydrogen bonded to O2P and O1P, respectively, of m5C-8. These interactions bring the first half of the loop L1 close to chain B of DNA while the second half of the long loop interacts with the phosphate group of G9 and T30 via hydrogen bonds from Ser116 and Lys119, respectively, in which, the latter bring the C-terminal end of this loop to chain C. This DNA induced movement of loop L1 is shown in Figure 5-8(inset) with a C_α RMSD of 3.4Å between amino acid Arg111 to Lys119 of the X-ray and NMR (Wakefield *et al.*, 1999) determined MBD structures. In addition, the side-chains of Lys109 of β1 and Ser134 of L3 also form hydrogen bonds with A7 and m5C-33, respectively (Figure 5-8).

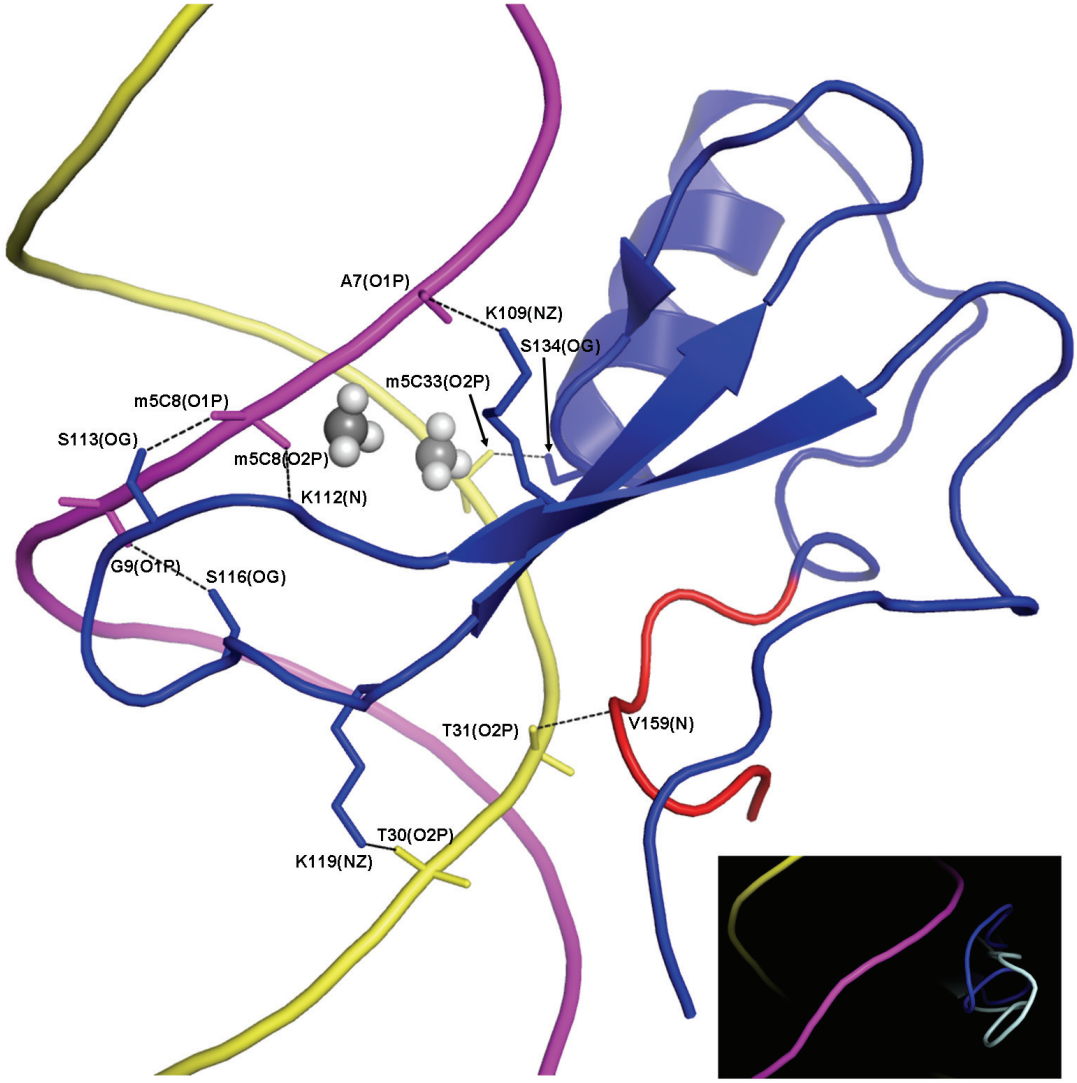


Figure 5-8 Hydrogen bonds between the DNA phosphate backbone and MeCP2 MBD

The cartoon represents the complex of MeCP2 MBD-DNA with the protein side/main-chain (sticks) hydrogen bonded to the DNA phosphate groups. All atoms involved in direct hydrogen bonds (black dash lines) are labelled (Table 5-7). Chains B and C of the DNA are coloured in purple and yellow, respectively. The methyl groups are indicated by grey spheres. Four of the seven hydrogen bonds position L1 in the major groove of the DNA. A hydrogen bond is also formed by the amide NH group of Val159 in the middle of the tandem Asx-ST-motif (red). The inset shows the DNA induced movement of Loop L1 of the X-ray structure (blue) compared to unliganded MeCP2 MBD (white loop). The RMSD calculated using Cα from residues 111-119 is 3.4Å.

The tandem Asx-ST-motif of the MeCP2 MBD also interacts with the DNA via a hydrogen bond between Val159(N) of the Asx-ST-motif and the phosphate oxygen of T31 (base-paired with A11) at the start of the AT run (^{11}AATT^{14}) (Table 5-7, Figure 5-8). Runs of 4-6 consecutive A/T bases adjacent to the methyl-CpG dinucleotides are required to maximise the binding of MeCP2 and methylated DNA (Klose *et al.*, 2005). This hydrogen bond (3.19 Å) is the only the direct interaction between the Asx-ST-motif and the AT run of the DNA, and may possibly play a major role in maintaining a stable Asx-ST-motif-DNA interaction. Alteration of amino acids composed of Asx-ST-motif which disrupt this hydrogen bonding could significantly reduce MeCP2-methylated DNA binding. This will be discussed in Chapter 6.

Table 5-7 Direct hydrogen bonds between MeCP2 MBD and the DNA phosphate backbone

Residues(atom)	Phosphate group(atom)	Distances (Å)
K109 (NZ)	A7(O1P)	3.1
K112(N)	m5C8(O2P)	2.9
S113(OG)	m5C8(O1P)	2.6
S116(OG)	G9(O1P)	3.4
K119(NZ)	T30(O2P)	2.9
S134(OG)	m5C33(O2P)	2.5
V136 (CB)	m5C33 (O2P)	3.5
V159(N)	T31(O2P)	3.2

All distances are below 3.6Å

The hydrogen bond that bridges Ser134 (loop L3) and Val136 (helix α1) to the oxygen phosphate of m5C-33 connects the N-terminal region of α-helix close to the DNA-protein contact interface. Water mediated contacts also contribute to the stabilisation of this region. Three water molecules (W5, W10 and W42) form hydrogen bond bridges between Ser134 and Val136 and the phosphate groups of m5C-33. W42 bridges the interaction between V136(CG2) to O2P of m5C-33. W5 coordinates the interactions of side-chain Ser134 and Val136 to atom O1P of m5C-33. Interestingly, W10 which is also involved in CH…O hydrogen bonds to methyl groups of m5C-33 and phosphate oxygen contacts is also hydrogen bonded with Ser134.

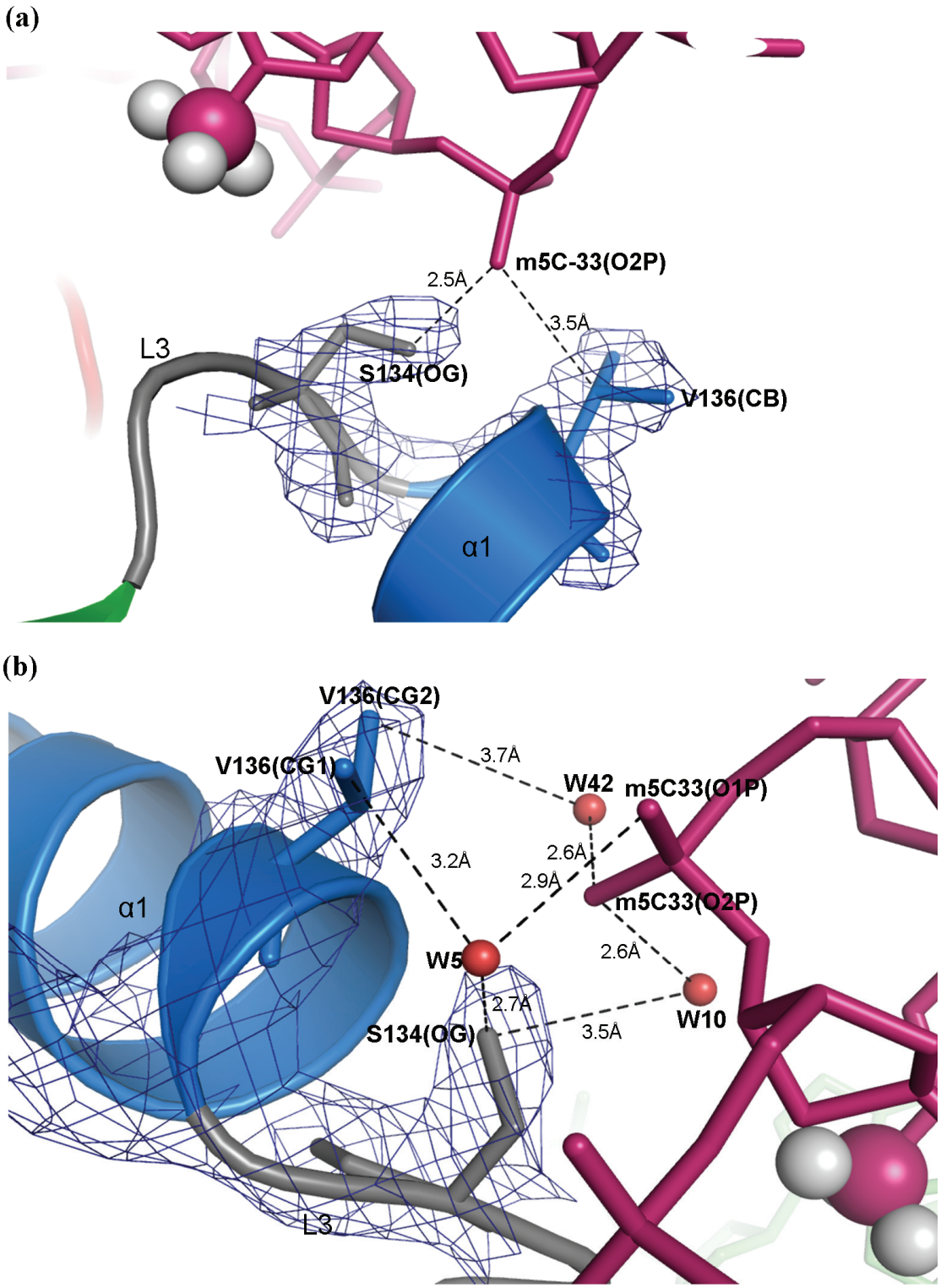


Figure 5-9 MBD alpha-helix 1 and DNA backbone interactions

(a) The only α-helix in the MBD domain is brought close to the DNA backbone (chain B) via hydrogen bonds formed between Ser134(OG) (on loop L3) and Val136(CB) (on α-1) with atom O2P of m5C-33. Loop L3 is coloured in gray and α-helix in blue. Methyl group of m5C-33 is shown as spheres.

(b) Three water molecules; W5, W10, and W42 connecting the Ser134 and Val136 to the phosphate oxygen of m5C-33 through several water mediated interactions. All water mediated interactions in this short loop are less than 4.0 Å.

5.3.5 DNA GEOMETRY

The primary sequence of the double stranded DNA influences the DNA groove width, helical twist, bending and mechanical rigidity or resistance to bending. These special features help other molecules such as repressors/activators to read and recognise their cognate DNA binding sequence. Early reports argued that the MBD domain of MeCP2 binds to methylated DNA regardless of their flanking sequences. The requirement of an AT run adjacent to the methyl-CpG to maximise the MBD binding however suggests that MeCP2 target genes are sequence specific (Klose *et al.*, 2005). The 20 mer *BDNF* gene fragment used in this study contains a central methyl-CpG and a run of AT bases which may exhibit the required DNA geometry. The A/T track DNA can be defined as a stretch of DNA containing four to six consecutive A/T bp. Depending on the DNA sequence, A/T track DNA is rigid and straight with a very narrow minor groove and a high degree of propeller twist (El Hassan and Calladine, 1996; Mack *et al.*, 2001; Nelson *et al.*, 1987; Prive *et al.*, 1987; Stefl *et al.*, 2004; Wing *et al.*, 1980). In order to study these properties, the DNA geometry of the X-ray structure was analysed using 3DNA (Lu and Olson, 2003).

5.3.5.1 Geometrical description of dinucleotide steps

The geometrical descriptions of DNA used in this section are simplified in a schematic diagram as shown in Figure 5-10. For the dinucleotide step parameters (Figure 5-10a), each block represents a base-pair with the minor groove sites coloured grey. The 5' to 3' direction of the strands is indicated (El Hassan and Calladine, 1996). The arrows give the directions of the rotations and translation of the upper, relative to the lower block, which correspond to a positive value of helical twist, roll and slide. To illustrate the base pair parameter of propeller twist as shown in Figure 5-10b, one base (represented by a rectangular block) is rotated relative to its pair in a left-handed sense (negative propeller twist) about the common long axis of the base-pair (El Hassan and Calladine, 1996).

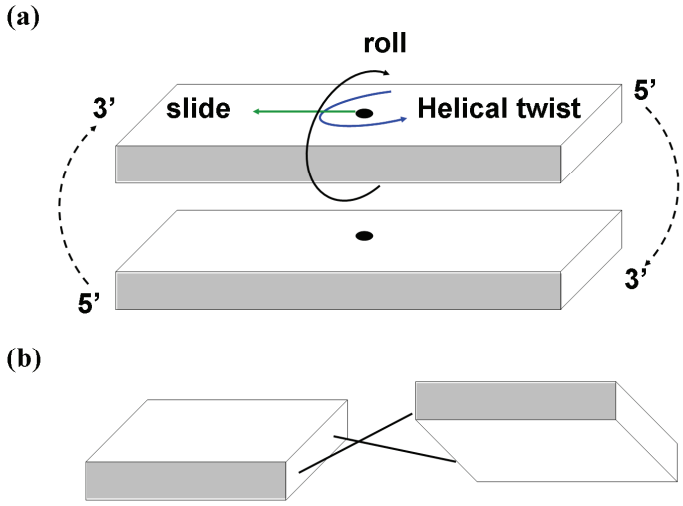


Figure 5-10 Schematic representations of (a) dinucleotide step and (b) base-pair parameters

5.3.5.2 Overall DNA geometry

A schematic drawing of the 20 mer *BDNF* helix is presented in Figure 5-11. The X-ray structure of the methylated *BDNF* fragment is a Watson-Crick right handed B-DNA helix with an average rise per residue of 3.3 Å (Table 5-8) which is shorter than standard B-DNA (3.4 Å) This is consistent with the observation of a cluster of strong DNA reflections at resolution close to 3.3 Å (Figure 4-9, Chapter 4) corresponding to base pair stacking in the DNA molecules. The 20 bp DNA helix composed of two complete turns, with ~10 bp per turn.

The minor and major groove widths were measured considering the directions of sugar-phosphate backbones. A useful measurement of the phosphate-phosphate distance across the major and minor groove opening is made by subtracting 5.8Å (the van der Waals radii of two phosphate groups) from the calculated values (El Hassan and Calladine, 1998). This value is not subtracted from the calculated phosphate-phosphate distances presented in Table 5-8. As shown in Figure 5-12, the narrowest minor groove of the DNA is located at the central run of AT bases (^{11}AATT^{14} paired with ^{28}AATT^{31}) with an average of 9.1 Å with the shortest distance of 8.6Å for the step A12T13/A29T30. On the other hand, the minor groove composed of ^{14}TCTTC^{18} paired with ^{24}GAAGA^{28} also displayed the characteristic of narrow minor groove of A/T track DNA with an averaged groove width of 9.8 Å (Table 5-8) which is narrower than the standard B-DNA (11.5 Å). In fact, 50% (^{11}AATTCTTCTA^{20}) of the

BDNF sequence used in this study can be considered as A/T track although intercalated by base pairs C15/G27 and C18/G24. Another region of the helix that shows significant narrow minor groove features is ^{6}AAm5C^{8} with an average minor groove width of 9.7 Å even though it consists only of two consecutive A/T base pairs.

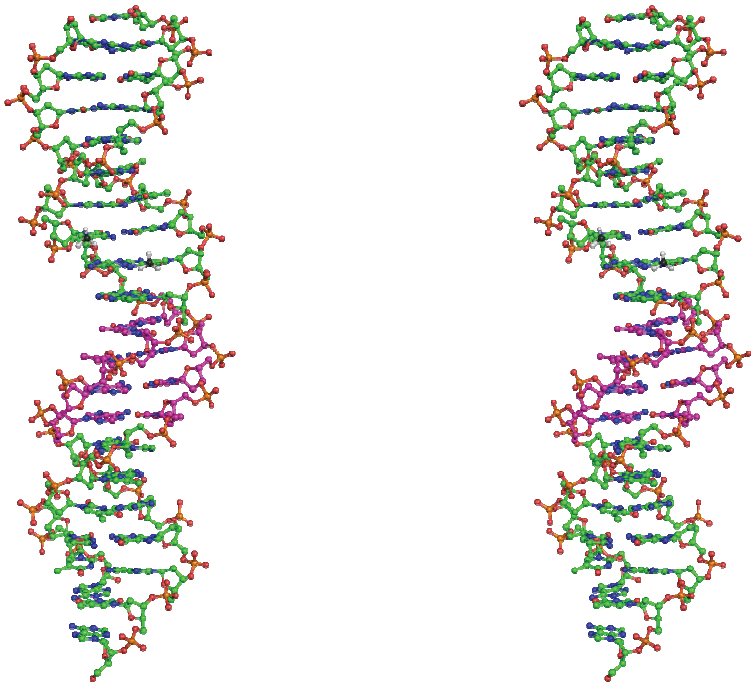


Figure 5-11 Skeletal stereo drawing of X-ray determined 20 mer *BDNF* fragment
The double-stranded DNA adopted a Watson-Crick right-handed B-DNA conformation and consists of 2 complete turns. This drawing is presented in the same orientation as in Figure 5-2b, where the numbering can be referred. The AT run of this DNA fragment is highlighted in red.

Table 5-8 Local helix parameters

Step	Minor groove/Å		Major groove/Å		Helix twist/°	Rise per base/Å
	BDNF	**Std B-DNA**	***BDNF***	**Std B-DNA**		
C2T3/A39G40	-	-	-	-	29.0	3.0
T3G4/C38A39	-	-	-	-	34.5	3.2
G4G5/C37C38	12.3	11.5	17.6	17.2	33.1	3.4
G5A6/T36C37	10.8	11.5	17.3	17.2	40.1	3.4
A6A7/T35T36	9.6	11.5	16.7	17.2	36.8	3.3
A7m5C8/G34T35	9.8	11.5	17.6	17.2	37.4	3.4
m5C8G9/m5C33G34	11.2	11.5	19.0	17.2	42.3	3.4
G9G10/C32m5C33	12.3	11.5	17.8	17.2	27.6	3.0
G10A11/T31C32	11.5	11.5	16.6	17.2	40.4	3.5
A11A12/T30T31	9.7	11.5	18.7	17.2	38.2	3.2
A12T13/A29T30	8.6	11.5	17.6	17.2	32.4	3.4
T13T14/A28A29	9.0	11.5	20.0	17.2	35.7	3.2
T14C15/G27A28	9.0	11.5	19.5	17.2	39.3	3.5
C15T16/A26G27	9.2	11.5	17.4	17.2	28.7	3.3
T16T17/A25A26	10.2	11.5	17.7	17.2	41.9	3.1
T17C18/G24A25	10.0	11.5	16.6	17.2	42.7	3.4
C18T19/A23G24	-	-	-	-	29.8	3.5
T19A20/T22A23	-	-	-	-	45.7	3.2
Average	**10.2**	**11.5**	**17.9**	**17.2**	**36.4**	**3.3**

All major and minor groove widths were measured using 3DNA (Lu and Olson, 2003). Direct P-P distances which take into account the directions of the sugar-phosphate backbones. Subtract 5.8 Angstrom from the values to take account of the vdw radii of the phosphate groups (El Hassan and Calladine, 1998).

By examining the major groove widths across the 20 bp DNA, the widest inter-phosphate distance across the major groove is located at step T13T14/A28A29 (20Å) which is coincidently one step further than the narrowest phosphate-phosphate distance for the minor groove located at step A12T13/A29T30 (8.6Å) (Figure 5-12b). Overall, the DNA helix is bent due to a kink at step G10A11/T31C32 which is immediately after the methyl-CpG dinucleotides. Figure 5-12a shows an overlay of standard B-DNA and the X-ray DNA in complex with the MeCP2 MBD domain. The X-ray DNA is bent to left at the beginning the of AT run. The RMSD fit of the four phosphorous atoms corresponding to the symmetrical methyl-CpG dinucleotides is 0.44 Å. In summary, the inter-phosphate distances in the X-ray structure showed a

minor groove that is significantly narrower than an ideal B-DNA minor groove (Figure 5-12b). Coincidently, the inter-phosphate distances of the major groove at this region are significantly wider than other regions of the DNA. This observation leads to the speculation that the DNA bend due to the AATT run is required to accommodate the stabilising interaction with the Asx-ST-motif in intact MeCP2.

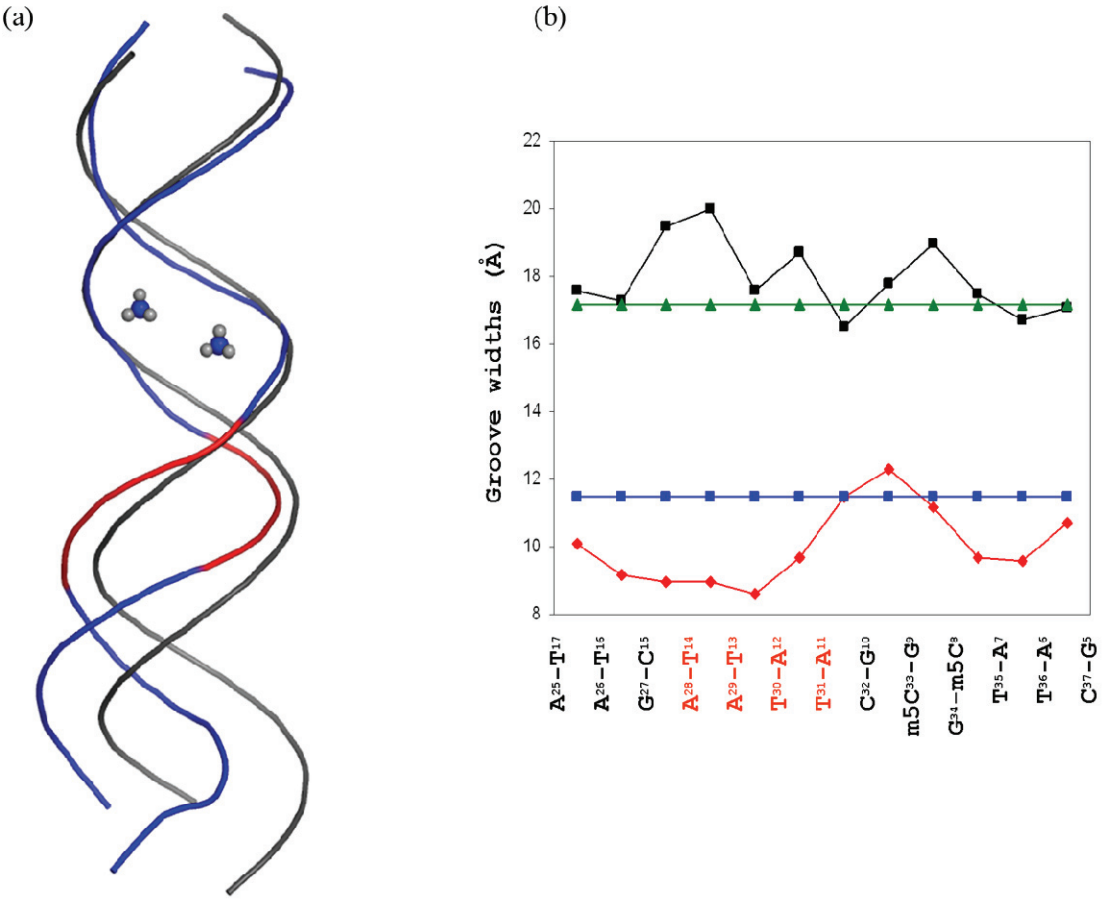


Figure 5-12: DNA minor and major groove widths
(a) Overlay of idealized B-DNA (grey) and X-ray DNA (blue and red) in complex with MeCP2 MBD domain. The rmsd fit of the four phosphorous atoms corresponding to the mCpG dinucleotides is 0.4 Å. The AT run is shown in red and the methyl groups of the mCpG dincleotides are shown as spheres.
(b) The minor groove width (red) and the major groove width (black) were calculated using the program 3DNA (El Hassan and Calladine, 1998; Lu and Olson, 2003). Groove widths for standard B-DNA are shown as a blue line (minor groove) and a green line (major groove). The minor groove at the AT run (5'^{11}AATT^{14}3') narrows significantly.

5.3.5.3 High degree of propeller twists

One of the most obvious characteristics of the A/T track DNA is that the bases of a base pair are not co-planar to a common axis. This is attributed to a principal characteristic of A/T track: high degree of propeller twist (Drew *et al.*, 1981; El Hassan and Calladine, 1996; Wing *et al.*, 1980). A definition for propeller twist is given in the previous section (Figure 5-10). As shown in Table 5-9 and Figure 5-13, propeller twists range from -2.4 to -20.3˚ with an average of -10.6˚. As a short A/T tract segment, the base pairs ^{11}AATTCTT^{17} exhibit a significantly high degree of propeller twist with an average of -15.5˚ even though intercalated by C15/G27 bp.

Table 5-9 Propeller twists of *BDNF* and standard B-DNA in degree

Base pair	***BDNF*** **Propeller twist/˚**	**Std B DNA** **Propeller twist/˚**
C2/G40	-5.5	-1.3
T3/A39	-3.3	-1.2
G4/C38	-7.6	-1.3
G5/C37	-10.6	-1.3
A6/T36	-9.8	-1.2
A7/T35	-8.8	-1.2
m5C8/G34	-4.6	-1.3
G9/m5C33	-7.5	-1.3
G10/C32	-2.4	-1.3
A11/T31	-19.7	-1.2
A12/T30	-14.2	-1.2
T13/A29	-15.4	-1.2
T14/A28	-15.0	-1.2
C15/G27	-14.2	-1.3
T16/A26	-10.4	-1.2
T17/A25	-19.4	-1.2
C18/G24	-5.4	-1.3
T19/A23	-20.3	-1.2
A20/T22	-8.1	-1.2
Mean	**-10.6**	**-1.2**

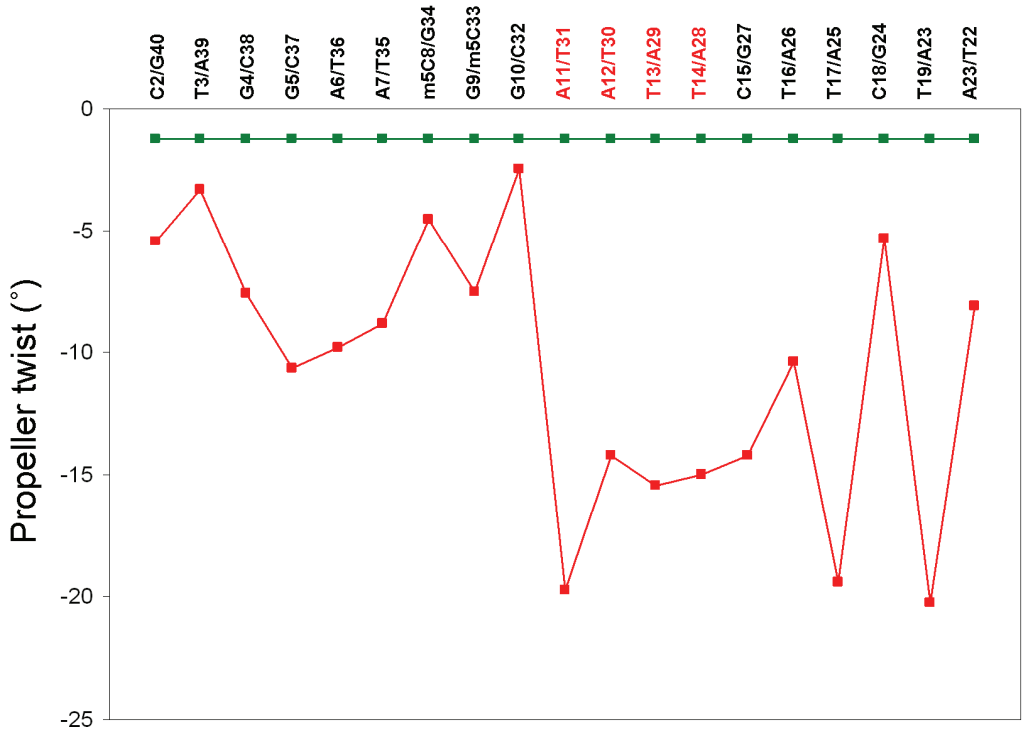


Figure 5-13 Propeller twists of *BDNF* sequence
A definition for propeller twist is given in the text. Plots refer to *BDNF* (red) and standard B-DNA (green) propeller twists in degree (°). Minus signs indicate left handed twists.

The high degree of propeller twist imposes two major effects on the A/T track DNA structure and improves overall stability of the DNA double helix. Twisting of bases within complementary base pairs creates an additional cross-strand diagonal hydrogen bond which helps purine-purine stacking interactions (Figure 5-14). In addition to the conventional Watson-Crick A/T base pairs, these cross-strand bifurcated hydrogen bonds can be seen connecting purine N6 to pyrimidine O4 in a diagonal position of the opposite strand (N6 of A11 to O4 of T30 and N6 of A28 to O4 of T13). Because of the high degree of propeller twist, the bases at the major groove site of the AT run are forced to point towards the 3' end of the self-strand. This pushes the N6 of A11 towards N4 of T30 which creates a non-Watson-crick hydrogen bond diagonally across the major groove (Figure 5-14c). Similarly, the amino group (N6) of A28 functions as a bifurcated proton donor to two carbonyl oxygen located on the opposite strand. Moreover, each carbonyl group of T13 and T30 can accept protons from N6 of two purine bases. Therefore, the AA/TT base pair can adopt a zig-zag system as shown in Figure 5-14d but this cannot happen in a mixed sequence such as AT/TA or GT/AC.

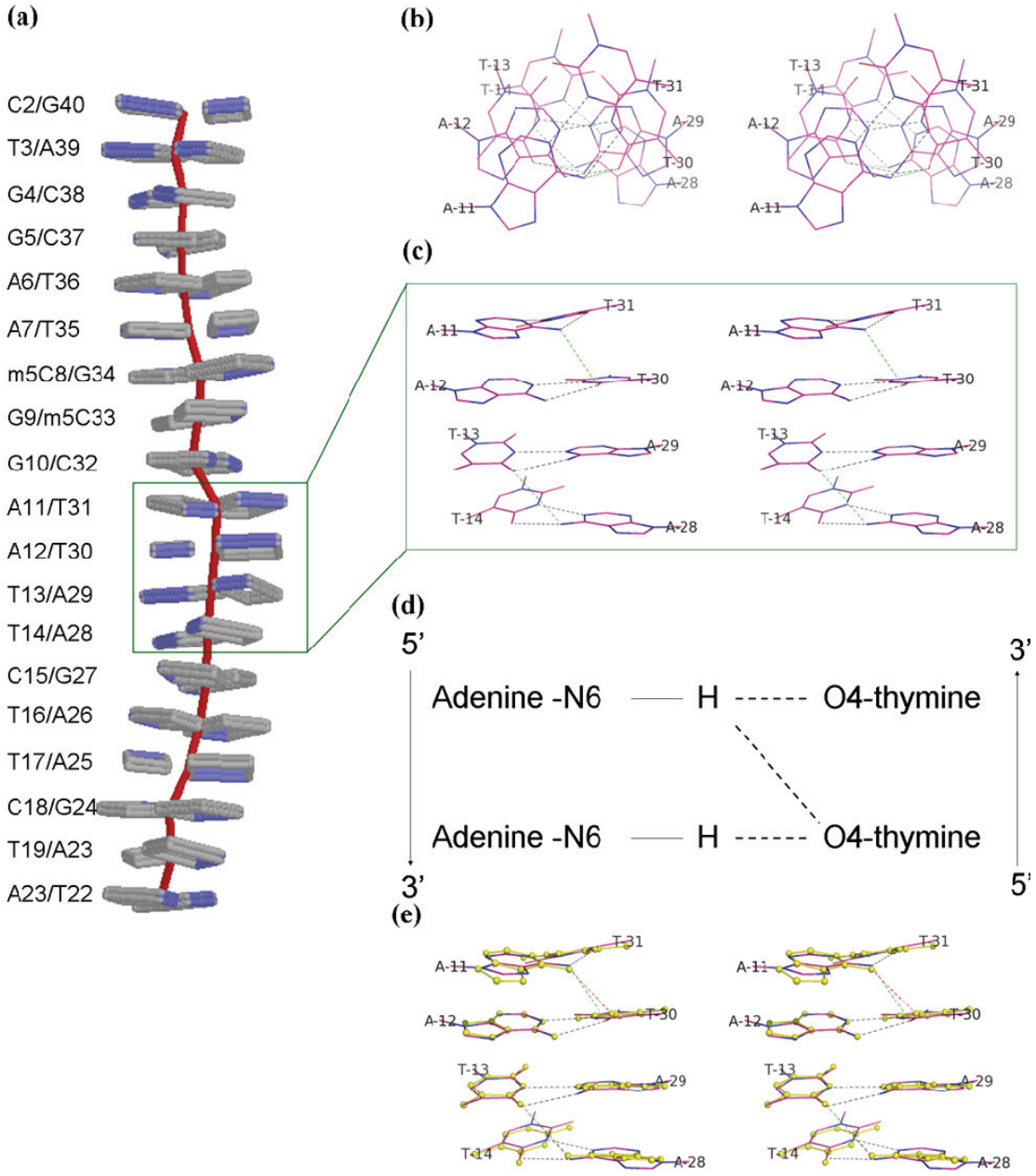


Figure 5-14 High degree of propeller twist at AT run base pair

(a) Each base is represented by a rectangular blocks. Big and small boxes represent purine and pyrimidine, respectively, with the minor groove sites coloured in blue. The middle red line represents the centre of the helix and the bases are numbered as in Figure 5-2. The AT run is within the green box, which has been enlarged (different orientation) to give (c).

(b) Stereo view of ^{11}AATT^{14}/^{28}AATT^{31} base steps looking down the helix from A11/T31 bp. The black and green dashes represent the conventional Watson-Crick hydrogen bonds between base pair and the non-Watson-Crick hydrogen bond diagonal across the major groove of the AT run, respectively. All bases are labelled appropriately.

(c) Stereo view of ^{11}AATT^{14}/^{28}AATT^{31} base steps looking at the opening of the major groove at the AT run.

(d) A zig-zag conformation of base pair step resulted from cross-strand non-Watson-Crick hydrogen bond. The direction of the DNA strands is indicated from 5' to 3'.

(e) Stereo view of overlay of the X-ray *BDNF* (red) and CGCGAATTCGCG (Wing *et al.*, 1980) (yellow spheres with sticks). The RMSD fit of the 60 ring atoms from the eight bases is 0.267Å. The orientation and number of the bases are identical as in (c)

As a result, no additional hydrogen bond was observed at step A12T13/A29T30. Although the formation of a cross-strand hydrogen bond has not been terminated after the ^{11}AATT^{14} bp, it has been interrupted by insertion of C15/G27. The cross-strand hydrogen bond at step C15T16/A26G27 does not contribute to the zigzag configuration because a bifurcated proton donor is absent at position C2 of G27 but the zigzag system is resumed at bp step T16T17/A25A26. The result is in agreement with other A/T track double-helix structures reported elsewhere (Fratini *et al.*, 1982; Nelson *et al.*, 1987; Wing *et al.*, 1980). Overlay of the X-ray *BDNF* and a helix composed of CGCGAATTCGCG (Wing *et al.*, 1980) shows strikingly similar features such as high degree of propeller twist and narrow minor groove of A/T track bp with an RMSD fit of 0.267Å for all 60 atoms of the A/T bases (Figure 5-14d).

Another important stabilising force of the A/T track DNA is the base stacking, especially purine-purine stacking, and this can result from a high degree of propeller twist which rotates bases within base pairs along their longitudinal axis, resulting in more interactions between neighbouring bases (Nelson *et al.*, 1987). As shown in Figure 5-14b, the adenines are stacked on top of each other with their 6-membered rings heavily overlapped whereas the thymine bases are barely overlapped and only make weak van der Waal's carbon-carbon contacts between the C5 methyl group and C6 of the subsequent thymine. The base-base compact stacking effect is not observed in GpG or methyl-CpG of the X-ray studied *BDNF* fragment. As a result of this compact stacking arrangement, the DNA sugar-phosphate backbones are closer compared with other regions of the DNA; this effectively narrows the minor groove of the AT run (Figure 5-12).

5.3.5.4 DNA hydration

Owing to the limited resolution (2.5 Å) of this X-ray structure, only 47 water molecules were located in the asymmetric unit. Along the 20 bp double helix *BDNF* fragment, the most humid region is located at the major groove where the protein and DNA interactions take place. The water molecules at this protein-DNA contact interface mediate the interactions between the methyl-recognition residues of the MeCP2 MBD domain and the methyl groups of methyl-CpG dinucleotides (see section 5.3.4.1). Some water molecules play space filling roles and stabilising the protein-DNA contacts. Other water molecules are positioned around the DNA helix.

It has been hypothesised that the high degree of propeller twist, which subsequently causes a narrow minor groove, might be stabilised by a spine of hydration in the AT rich region (Fratini *et al.*, 1982). These spines of hydration are believed to be required to preserve the integrity of the B-DNA helix. The disposition of water molecules around the AT run is shown in Figure 5-15a. A true spine of water molecules has not been observed in the X-ray *BDNF* at this resolution. There are however three structured water molecules that form a "short spine of hydration"; they run down the wall of the minor groove composed of the AT run. As seen in Figure 5-15b, the O2 atom of Thymine and N3 of Adenine (or Thymine O2) of adjacent base pairs are brought into a closer proximity by water bridging. On the minor groove site of the AT run, the bases are oriented towards the 5' direction on its strand in a way that the displaced O2 of Thymine or N3 of Adenine can be hydrogen bonded to the water molecules. The distances of these water bridges are tabulated in Table 5-10. This pattern is the same as was described by Prive and coworkers (1987): the pyrimidine O2 or purine N3 of base n is water bridged to O4'atom of deoxyribose (n+1) (Prive *et al.*, 1987). This configuration has been seen in diverse examples of A/T track DNA (Mack *et al.*, 2001). Because the sugar-phosphate strands are brought into a close proximity (consequence of narrow minor groove), the water molecules also connect the deoxyribose 5-membered ring at O4'. As shown in Figure 5-15b, two (W30 and W32) of these water molecules are bridged tetrahedrally to two base atoms and two sugar atoms. The unfilled tetrahedral coordination of W18 was due to a sharp helical rise at step G10A11/T31C32 (Table 5-8), which is coincident with the kink of the DNA helix, making the deoxyribose O4' of C32 unreachable by W18 (represented by red dashes in Figure 5-15a and b).

In the A/T track DNA, atoms N6 of Adenine and O4 of Thymine are brought into a close proximity in the major groove by the non-Watson-Crick cross-strand hydrogen bonds (Figure 5-15). Together with the cross-strand water bridges in the minor groove, these hydrogen bonds exert forces to twist the A/T base pair in an opposite directions. This effect significantly enhances base-base (particularly purine-purine) compact stacking. This brings the adjacent base pairs into close proximity. They are also stabilised by a well coordinated water bridging system in the minor groove.

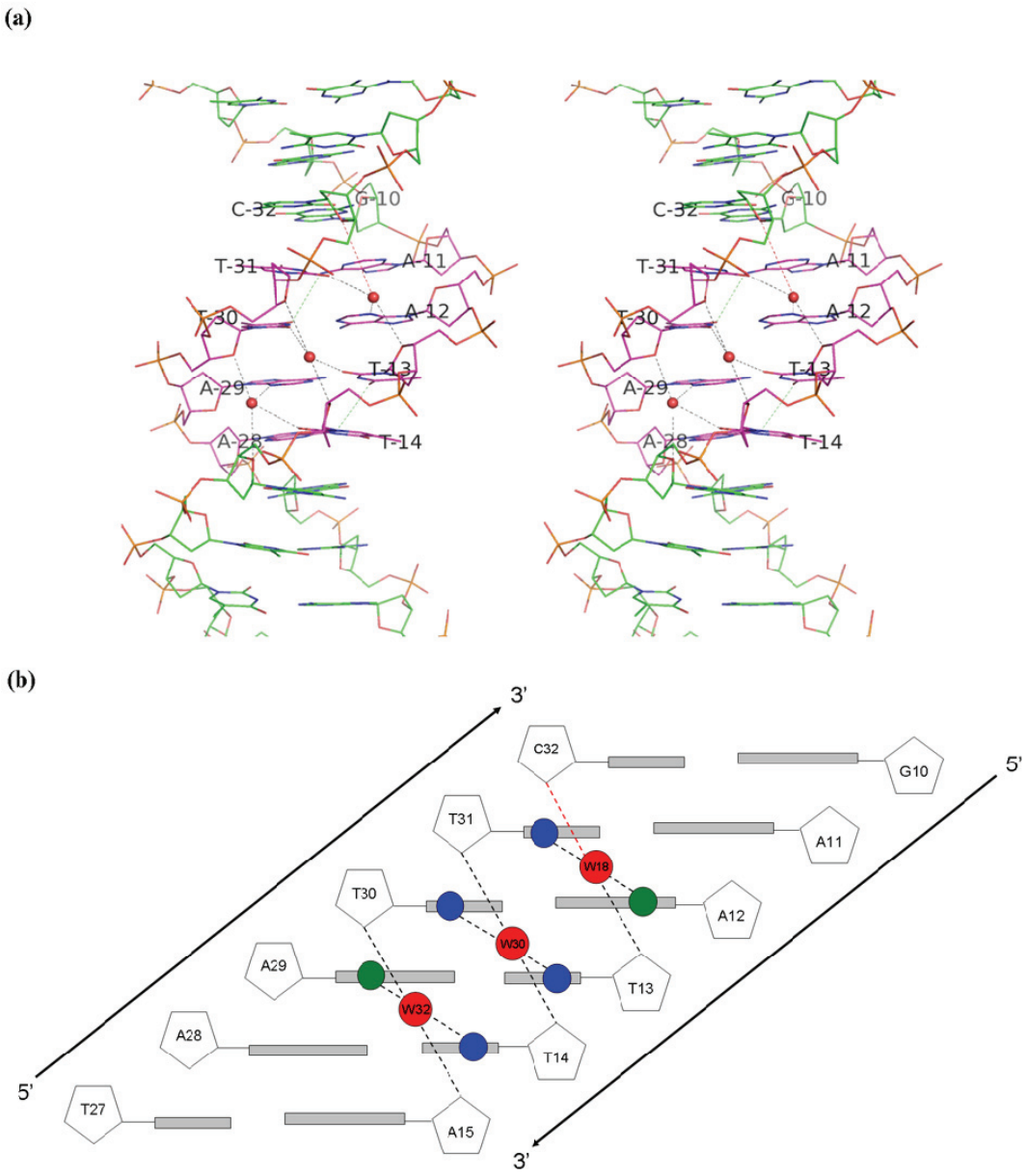


Figure 5-15 Hydration at the AT run

(a) Stereo view looking at the minor groove of the AT run. Running down the minor groove are W18, W30 and W32 with each water molecule (except W18) coordinating tetrahedrally to four atoms from Adenine or Thymine bases and their deoxyribose ring. Black dashes represent water bridged hydrogen bonds and red dashes indicate a potential hydrogen bond to the atom O4' of deoxyribose C32 if the T31C32 step is undistorted. The central A/T bases are labelled appropriately. The distances for the water mediated hydrogen bonds are listed in Table 5-10.

(b) Schematic diagram showing the water bridging system in the AT run. Each water molecule is hydrogen bonded to two bases and two deoxyriboses. The long and short horizontal grey bars represent purine and pyrimidine, respectively, and the five-membered-ring indicates the deoxyribose. Blue and green spheres represent atom O2 of Thymine and N3 of Adenine, respectively. The direction of each DNA strand is indicated. The red dashed line indicates a distance above 4Å from W18 to the deoxyribose O4' of C32 due to a sharp helical rise at step G10A11/T31C32.

Table 5-10 Water bridges at the AT run

Water molecule	nucleotide	Atom	Distance (Å)
W18	T31	O2	2.9
	A12	N3	2.9
	T13	Sugar O4’	3.1
W30	T30	O2	3.1
	T13	O2	2.7
	T14	Sugar O4’	2.9
	T31	Sugar O4’	3.4
W32	A29	N3	2.8
	T14	O2	3.0
	T30	Sugar O4’	2.9
	C15	Sugar O4’	3.3

5.4 SUMMARY

This chapter addresses the details of molecular interactions between the MBD domain and the methylated DNA. The X-ray structure reveals that the recognition of methyl-CpG by the MBD domain does indeed depend upon a modified hydration pattern in the DNA major groove. The binding specificity of methylated DNA has been reported to depend on hydrophobic interactions between the MBD domain and the methyl-CpG (Ohki *et al.*, 2001). However, the X-ray structure in this study shows that the methyl groups make contact with a predominantly hydrophilic surface that consists of several water molecules. The structural analysis also demonstrated that T158 and R106 (two of the four most frequent hotspots RTT mutations) play major roles in stabilising the tandem Asx-ST motif in MeCP2. Additionally, the AT track DNA in the X-ray structure displays unique features such as narrow minor groove and high degree of propeller twist. A short spine of water molecules has been observed running down the wall of minor groove of AT run. It is however remains elusive how these AT tract features enhance MeCP2 binding.

CHAPTER 6. MUTAGENESIS STUDIES

6.1 INTRODUCTION

Mutational analysis is a powerful tool to examine a hypothesis by changing amino acids in a protein sequence. Suspected critical residues in a protein can be substituted with other amino acids by altering the protein encoding DNA sequence. This can be achieved by site-directed mutagenesis. The amino acid substitution in site directed mutagenesis can be done by annealing a mutagenesis primer to a template followed by complementary strand synthesis using DNA polymerase. Traditionally, single-stranded DNA of M13 was used in the following way: a primer containing the desired nucleotide substitution was annealed to the template and elongated using DNA polymerase, ligated with DNA ligase to seal the nick. The mismatched duplex was then transformed into the *E. coli* expression system (Kunkel, 1985; Sugimoto *et al.*, 1989; Taylor *et al.*, 1985; Vandeyar *et al.*, 1988). Some bacteria repair the mismatch to give the desired point mutation whereas some clones retain their original sequence. Due to the inefficiency of this method, site directed mutagenesis using double stranded DNA was invented (Braman *et al.*, 1996). In general, a pair of complementary mutagenic primers usually between 25 and 45 bp is designed to have the altered sequence. After the PCR amplification, the parental plasmids can be easily digested with *Dpn* I which is an endonuclease that is specific for methylated and hemimethylated DNA. The mutant plasmids are then transformed into the desired expression system. This method is highly efficient and can be done within 2-3 days. The mutated gene within the mutant plasmids need to be verified with DNA sequencing before protein expression and purification. On the other hand, mutation of oligonucleotides is more straightforward. A new pair of oligonucleotides containing the desired nucleotide substitutions can be synthesised directly from a manufacturer.

As discuss in Chapter 5, certain residues within the MBD domain are crucial in maintaining the protein-DNA interaction. X-ray analysis reveals that residue T158 plays a pivotal role in shaping the Asx-ST motif which comprises 5 residues. Substitutions of T158 with other residues such as Ala, Met and Ser confirmed the importance of T158 in stabilising the Asx-ST motif. Gel shift analysis using mutant Y123F has been performed to assay the importance of the water bridging role of

Y123. This experiment highlighted the important role of the hydrated interface between the MBD domain and the DNA major groove. This chapter will also discuss the mutational analysis of the oligonucleotides of the *BDNF* sequence. It is speculated that the AT run close to the methyl-CpG plays a specific role in enhancing the MBD-DNA interactions (Klose *et al.*, 2005). X-ray structure analysis reveals that this proximal AT run connects to the Asx-ST motif via a hydrogen bond. Additionally, a distal AT run, which is located approximately 10 bp upstream from the methyl-CpG is speculated to form another interaction site between the AT hook of MeCP2 and the methylated DNA. Therefore, various constructs have been used to assay the DNA-protein interactions.

6.2 MATERIALS AND METHODS

Materials and methods used in protein purification and gel shift assay can be found in Chapter 3.

6.3 RESULTS AND DISCUSSION

6.3.1 Mutational studies of Threonine-158

As discussed in Chapter 5, the X-ray structure reveals that Thr158 occupies a pivotal position that coordinates an Asx turn and an ST-motif (Wan and Milner-White, 1999a; Wan and Milner-White, 1999b), which we subsequently named as the Asx-ST motif (Ho *et al.*, 2008). The T158M mutation is the most common missense mutation causing Rett syndrome. To test this, the effects of T158 mutations on DNA binding were assayed. As shown in Figure 6-1, T158M abolished DNA binding under the experimental conditions in this study. Substitution of Thr158 by smaller alanine side chain (T158A; also a RTT mutation) again strongly impaired DNA binding, arguing against a purely steric interference by the bulky methionine side chain. In contrast, T158S, which retains the hydroxyl group that is implicated in hydrogen bond stabilisation of the tandem Asx-ST motif, maintained substantial affinity for methylated DNA (Figure 6-1). The result also indicated that the methyl group of Ser is not required for the stabilisation of the Asx-ST motif.

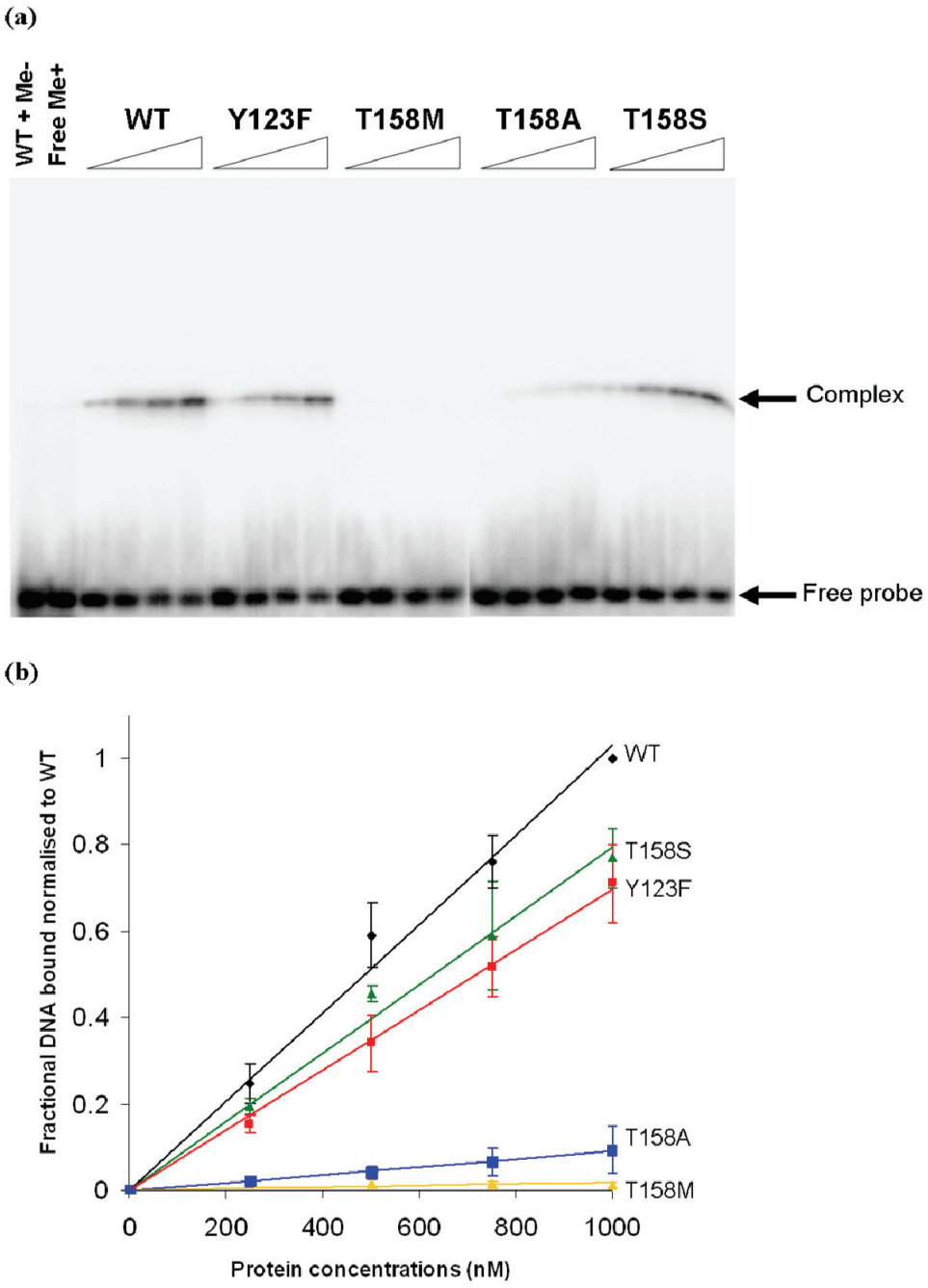


Figure 6-1 Mutagenesis confirms the importance of Y123 and T158 for the MBD binding to methylated DNA *in vitro*

(a) Electrophoretic mobility shift assays were performed with the 20 bp *BDNF* DNA sequence as probe in the presence of wild type (WT) or mutant forms of the 77-167 fragment of MeCP2. Binding to methylated DNA is disrupted by T158M and T158A mutations, whereas binding is minimally affected by the conservative T158S mutation. The Y123F mutation significantly reduces DNA binding.

(b) A plot of the fraction of labelled probe complexed at protein concentrations of 250nM, 500nM, 750nM and 1µM as measured by densitometry. Estimated standard deviations (shown as vertical bars) were calculated from measured intensities from three separate gels and scaled to the band from the probe complexed with 1µM wild type MBD which was taken as 100%. Plots refer to WT (black), T158A (blue), T158M (yellow), T158S (green: from two measurements) and Y123F (red).

Disruption of the Asx-ST motif as a result of the missense mutation T158M/A possibly impairs the two hydrogen bonds that connecting the guanidinium group of R106 to the main-chain carbonyl of T158 and V159 (see Figure 5-4a). This result is further supported by mutational analysis using R106W and T158M, where both mutants failed to recognise the methylated DNA (Yusufzai and Wolffe, 2000). The removal of the hydroxyl group of Tyr or Ser impairs the hydrogen bonds which functionally stabilise the ST-motif. It is also likely that interruption of the contacts between R106 and Asx-ST motif will destabilise the protein. In contrast to MeCP2, R17 of MBD1 (equivalent to R106 in MeCP2) makes contact with a β-turn composed of ^{63}DFKQ^{66}, in which, the mutation of K65A (K65 is equivalent to T158 in MeCP2) also abolished methylated DNA binding (Ohki *et al.*, 2001). This suggests that the β-turn of MBD1 MBD domain may play a similar role as the Asx-ST motif in maintaining the stability of the C-terminal end of the MBD domain. In general, the results of this study and others suggest the side-chain hydroxyl group of either Ser or Thr is critical to maintaining the structural role of the Asx-ST motif and the DNA selectivity of human MeCP2.

6.3.2 Mutational studies of Asp121 and Tyr123

The importance of water molecules in MBD recognition is supported by mutagenesis studies of MBDs from MeCP2 and MBD1. Y123 is of particularly interest, as its contact to the DNA is via two bridging hydrogen bonds from its hydroxyl group to W24 and W12 (Figure 5-5, Chapter 5). Figure 6-1b shows that the Y123F mutant lacks this hydroxyl group and has a reduced affinity (approximately 30% of the wild type protein) for methylated DNA. The loss of binding is therefore attributed solely to the interaction of the hydroxyl group with "structural waters" W24 and W12, because no other DNA contacts less than 4 Å are made by the Y123 side chain.

In a corresponding mutagenesis study of MBD1, reduced methyl-CpG binding caused by loss of the hydroxyl group of Y34 (Y34F is equivalent to Y123F in MeCP2; see Figure 5-1, Chapter 5) was explained by loss of a putative hydrogen bond between Y34 and the amino group of m5C (Ohki *et al.*, 2001). The X-ray structure of this study fails to support the equivalent Y123-m5C interaction in MeCP2, and a more likely explanation in the light of this closely analogous MeCP2 structure is that

hydrogen bonds to bridging structural water molecules have been lost in the Y123F mutant (Figure 5-5b, Chapter 5). Further evidence for the importance of this interaction comes from mutagenesis of mammalian MBD3, which does not normally bind specifically to the methylated DNA and has phenylalanine (F34 in MBD3) in place of tyrosine at the equivalent position (see Figure 5-1, Chapter 5). Interestingly, MBD3 can be converted into a methyl-CpG-binding protein by addition of a hydroxyl group via a substitution of F34Y (Fraga *et al.*, 2003). In addition to Y123, D121 is involved in water coordination, even though this residue also forms a direct CH…O interaction with the methyl group of m5C8 (Figure 5-5, Chapter 5) and two hydrogen bonds to R111. Mutagenesis of this aspartate in MeCP2 (D121A or D121C; Free et al., 2001) or the equivalent residue in MBD1 (D32A; Ohki et al., 2001) severely reduces binding to methylated DNA. The mutagenesis data derived from several members of the MBD protein family therefore support the structural evidence that water is a key feature for recognition of methyl-CpG by MBD domains.

The importance of water-mediated recognition mode established in this study is proposed to be conserved in other members of the MBD protein family. The X-ray structure of the MeCP2 MBD complexed with methylated DNA indeed resembles the solution structure of the MBD1 MBD-DNA complex structure reported by Ohki and coworkers (Ohki *et al.*, 2001) with an RMSD fit of 4.2 Å (Figure 5-3a, Chapter 5). It was previously deduced that the recognition of the cytosine methyl groups by MBD1 depends upon a hydrophobic patch consisting of 5 amino acids that are conserved among the MBD family (Ohki *et al.*, 2001). However, water molecules were not considered in the MBD1 structure determination, and it is possible that analysis of spin diffusion effects mediated by tightly bound waters will permit reconciliation of the NMR data with the hydrogen bond configuration reported in this study.

6.3.3 Mutational studies of the *BDNF* sequence

The promoter region of the *BDNF* gene was identified as an endogenous MeCP2 target (Chen *et al.*, 2003; Martinowich *et al.*, 2003). A methyl-CpG pair is located at position -100 bp upstream of the *BDNF* transcription initiation site, which is followed by two AT runs, namely a proximal AT run (composed of 4 bp) and distal AT run (composed of 8 bp), one and 10 bases, respectively, upstream from the methyl-CpG

(Figure 6-2a). The presence of an AT-run has been shown as a requirement to promote high affinity binding of MeCP2 (Klose *et al.*, 2005). The X-ray structure indicates that the proximal AT-run (11AATT14) of the *BDNF* sequence close to the methyl-CpG interacts with the Asx-ST motif of the MeCP2 MBD domain (Figure 5-4, Chapter 5). To test the importance of AT bases of *BNDF* adjacent to the methy-CpG, the AT runs were individually or simultaneously mutated and assayed on protein binding using band shift experiments. As shown in Figure 6-2b, the MBD domain (construct 77-167) binding was not affected by the mutations of either proximal or distal AT runs. Because the only interaction between the MBD and the AT run is through a hydrogen bond connecting the T31 phosphate group and Val159(N), the altered AT run may not disturb the DNA-protein interactions in the major groove.

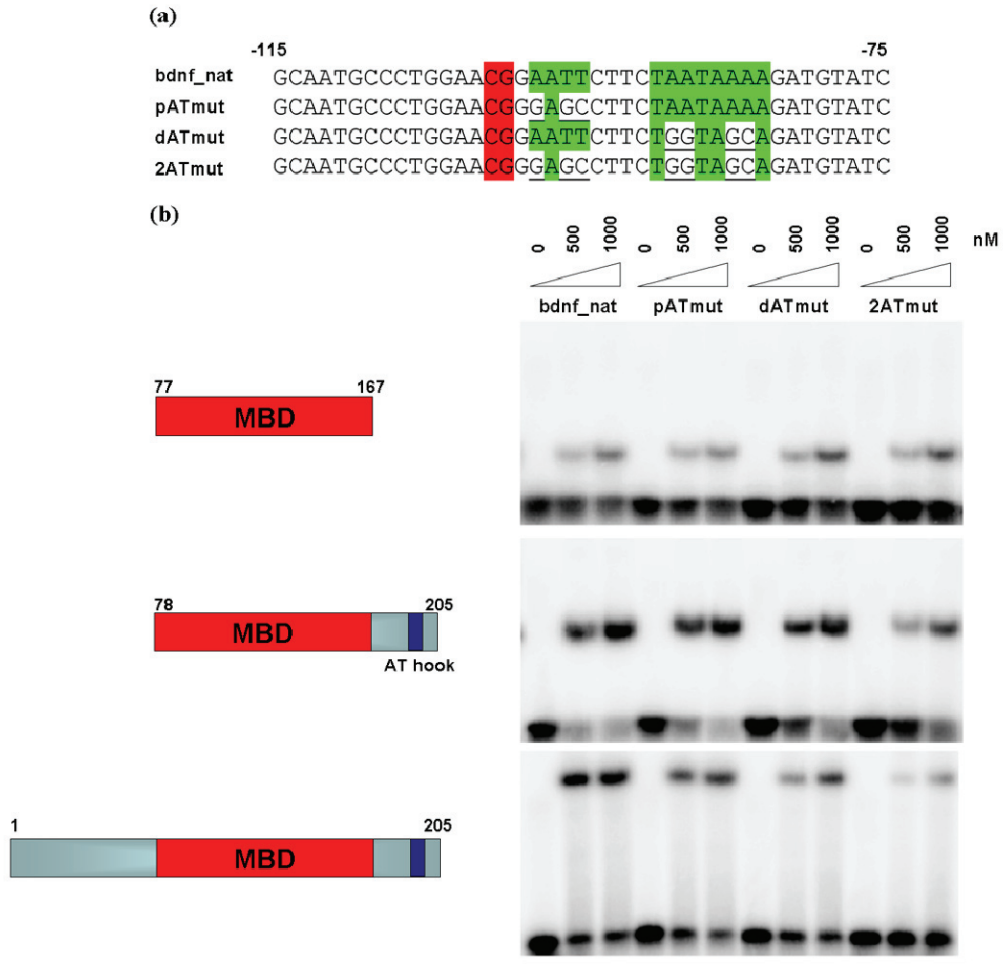


Figure 6-2 EMSA of proximal and distal AT mutations with MeCP2 constructs
(a) DNA probes used in the band shift assay. bdnf_nat: native *BDNF* sequence; pATmut: *BDNF* with proximal AT run mutated; dATmut: *BDNF* with distal AT run mutated; and 2ATmut: *BDNF* with both AT runs mutated. All DNA probes comprise of 41 bp. (b) Band shift assays were performed using three MeCP2 constructs; 77-167, 78-205, 1-205, 77-167(T158M) and 78-205(T158M).

Both longer constructs (78-205 and 1-205) show a significantly reduced binding when both AT runs were simultaneously mutated but not individually (Figure 6-2). The AT hook has been shown to interact with DNA bases in the minor groove of the AT tract DNA (Huth *et al.*, 1997). Close sequence similarity between the AT hook of HMGA1 and MeCP2 suggests that the AT hook of MeCP2 may interact with either proximal or distal AT runs. However, an AT hook mutation was shown to have little effect on methylated DNA binding (Klose *et al.*, 2005). Mutation of both AT runs simultaneously may alter the entire DNA geometry, including the major groove (and its hydration pattern), which drastically affects methyl group recognition.

6.4 SUMMARY

To test the pivotal role of T158 in the Asx-ST motif, T158 was mutated to Met and the mutation abolished DNA binding. Similarly, T158A also significantly reduced the DNA binding specificity arguing against a purely steric interference by the methionine side chain. Interestingly, retaining the hydroxyl group in the T158S mutation rescued affinity for methylated DNA. These mutational analyses support the argument that the hydrogen bonds connecting the hydroxyl group of either Ser or Thr at position 158 to the amine group of G161 and R162 are critical in stabilising the MBD domain of MeCP2. To test the importance of water molecules that mediate the methyl group recognition, Y123 was mutated to Phe and as expected, this alteration reduced DNA binding. Together with other reports (Fraga *et al.*, 2003; Ohki *et al.*, 2001), this result further strengthen the observed structural details. In the *BDNF* mutational studies, neither proximal nor distal AT runs to the methyl-CpG exert a significant effect on MBD binding specificity. Reduction was only observed when both AT runs close to the methylated site were mutated. This suggests that the overall DNA geometry was altered which subsequently affected the properties in the DNA major groove.

CHAPTER 7. CONCLUSIONS AND FUTURE DIRECTIONS

7.1 ARE THE AIMS ACHIEVED?

The major aim of this study was to investigate the molecular details of the MeCP2 MBD domain in complex with methylated DNA using X-ray crystallography. There are three basic questions which the X-ray structure of the DNA-protein complex can potentially address. (i) How is the methyl-CpG recognised by the MBD domain of MeCP2? (ii) What are the structural roles of the RTT disease hotspots such as R106 and T158 within the MBD domain? Are these critical residues interacting with the methylated DNA? and (iii) How does the AT run adjacent to the methyl-CpG enhance the MeCP2 MBD binding?

7.1.1 Novel X-ray structure of MeCP2 MBD complexed with methylated DNA

To investigate the structural details of methyl-CpG recognition by the MBD domain of MeCP2, a complex of the methylated DNA and MeCP2 was characterised (Chapter 3) and crystallised (Chapter 4). In collaboration with Dr. Robert Klose, the AT run next to the methyl-CpG, which maximises the protein binding, was taken into account in designing crystallisation trials. Cocrystals of MeCP2 MBD in complex with a 20 bp fragment of *BDNF* promoter which contains a central methy-CpG pair and an adjacent AT run was successfully crystallised. The best optimised crystal diffracted X-rays to a maximum resolution of 2.5Å using synchrotron radiation sources. The X-ray structure was solved using the selenium 'peak' data with the SAD method. In addition, the native X-ray structure and two iodinated derivatives were also solved and refined. The details structural analysis was carried out using the best refined model.

7.1.2 MeCP2 binding to DNA depends upon hydration at methyl-CpG

The molecular details reveal that the methyl groups of m5C in the major groove of the DNA are recognised by the MBD domain through water molecules (Ho *et al.*, 2008). This novel finding contradicts the observation using an NMR model (Ohki *et al.*,

2001), in which, the recognition is thought to depend on hydrophobic interaction between the methyl groups and a hydrophobic patch within the MBD domain. The hydration pattern around the 5'methyl-cytosine is modified in a way that the MBD can recognise the distinct hydration in the major groove of methylated DNA rather than methylation *per se*.

7.1.3 Thr158 and Arg106 are required to maintain Asx-ST motif

T158 is the major RTT mutation hotspot within the MBD domain. Structural investigation shows that this critical residue occupies a pivotal position in coordinating two consecutive turns, which is collectively named as the 'Asx-ST motif' (Ho *et al.*, 2008). The T158M mutation, the highest RTT missense mutation, abolished DNA binding in this study. Substituting T158 by the smaller Ala side chain again significantly reduced DNA binding, thus arguing against purely steric interference by the bulky methionine side chain. Interestingly, T158S, which retains the hydroxyl group that is implicated in hydrogen bond stabilisation of the Asx-ST motif, maintained a significant affinity for the methylated DNA. In addition, the tandem Asx-ST motif is further stabilised by hydrogen bonds to R106 via main chain carbonyl groups of T158 and V159. The structural role of R133, another major RTT missense hotspot, is to make direct contact with the guanine of methyl-CpG. Both R133 and R111 are hydrogen bonded to G34 and G9 bases, respectively, via their arginine fingers. Mutation of R133 that abolish DNA binding have been reported elsewhere (Free *et al.*, 2001; Ohki *et al.*, 2001; Yusufzai and Wolffe, 2000). Further biochemical or higher X-ray resolution is desired in order to address other RTT disease spots within the MBD domain. Moreover, the molecular details of RTT mutants might provide a fundamental knowledge in RTT therapeutic approaches.

7.1.4 How does AT run enhance MBD binding?

The AT run adjacent to the methyl-CpG displays some unique features of AT tract DNA such as a narrow minor groove, high degree of propeller twist, and a distinct pattern of hydration along the wall of minor groove. Although the Asx-ST motif of the MBD domain is hydrogen bonded to the AT run through a backbone phosphate group, one cannot draw a conclusion that this interaction is indeed required for high affinity binding. As a result, the contribution of the AT run in enhancing the MBD

binding remains unclear. Furthermore, the interaction occurs at the C-terminus in the X-ray structure. It is possible that other domains, for instance, the AT hook of MeCP2 might interact strongly with the AT run. To examine the structural role of an AT run next to methyl-CpG, a structure of a bigger MeCP2 construct including the AT hook domain in complex with methylated DNA is necessary. This large structure could possibly answer the question of how does the AT hook(s) bind into the minor groove of the DNA. The outcomes will provide invaluable information on MeCP2 binding partners and actual roles of MeCP2 as a transcription repressor or co-repressor.

7.2 PDB ACCESSION NUMBER

The atomic coordinates of the MeCP2 MBD domain cocrystallised with a 20 bp DNA fragment have been deposited with the Protein Data Bank (PDB) under accession number 3C2I.

7.3 PUBLICATION

Part of the crystallographic data (Chapter 4), structural details (Chapter 5) and mutational studies (Chapter 6) of this work has been published in Molecular Cell, volume 29, issue 4, pages 525-531 on 29th February 2008.

REFERENCES

Aapola, U., Kawasaki, K., Scott, H. S., Ollila, J., Vihinen, M., Heino, M., Shintani, A., Kawasaki, K., Minoshima, S., Krohn, K.*, et al.* (2000). Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. Genomics *65*, 293-298.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., and Terwilliger, T. C. (2002). PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr *58*, 1948-1954.

Amir, R., Dahle, E. J., Toriolo, D., and Zoghbi, H. Y. (2000). Candidate gene analysis in Rett syndrome and the identification of 21 SNPs in Xq. Am J Med Genet *90*, 69-71.

Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., and Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nat Genet *23*, 185-188.

Antequera, F., Macleod, D., and Bird, A. P. (1989). Specific protection of methylated CpGs in mammalian nuclei. Cell *58*, 509-517.

Banks, G. C., Mohr, B., and Reeves, R. (1999). The HMG-I(Y) A.T-hook peptide motif confers DNA-binding specificity to a structured chimeric protein. J Biol Chem *274*, 16536-16544.

Bartolomei, M. S., Zemel, S., and Tilghman, S. M. (1991). Parental imprinting of the mouse H19 gene. Nature *351*, 153-155.

Bell, A. C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482-485.

Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387-396.

Benning, M. M., Shim, H., Raushel, F. M., and Holden, H. M. (2001). High resolution X-ray structures of different metal-substituted forms of phosphotriesterase from *Pseudomonas* diminuta. Biochemistry *40*, 2712-2722.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R., and Schneider, B. (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J *63*, 751-759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J Mol Biol *203*, 971-983.

Bestor, T. H. (2000). The DNA methyltransferases of mammals. Hum Mol Genet *9*, 2395-2402.

Bestor, T. H., and Verdine, G. L. (1994). DNA methyltransferases. Curr Opin Cell Biol *6*, 380-389.

Bhattacharya, S. K., Ramchandani, S., Cervoni, N., and Szyf, M. (1999). A mammalian protein with specific demethylase activity for mCpG DNA. Nature *397*, 579-583.

Bienvenu, T., and Chelly, J. (2006). Molecular genetics of Rett syndrome: when DNA methylation goes unrecognized. Nat Rev Genet *7*, 415-426.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev *16*, 6-21.

Bird, A. (2007). Perceptions of epigenetics. Nature *447*, 396-398.

Bird, A. P., and Wolffe, A. P. (1999). Methylation-induced repression--belts, braces, and chromatin. Cell *99*, 451-454.

Birke, M., Schreiner, S., Garcia-Cuellar, M. P., Mahr, K., Titgemeyer, F., and Slany, R. K. (2002). The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. Nucleic Acids Res *30*, 958-965.

Blow, D. (2002). Outline of Crystallography for Biologists, 1 edn, (Oxford University Press).

Blundell, T. L., Johnson, L. (1976). Protein Crystallography, (Academic Press).

Boeke, J., Ammerpohl, O., Kegel, S., Moehren, U., and Renkawitz, R. (2000). The minimal repression domain of MBD2b overlaps with the methyl-CpG-binding domain and binds directly to Sin3A. J Biol Chem *275*, 34963-34967.

Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B., and Bestor, T. H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536-2539.

Braman, J., Papworth, C., and Greener, A. (1996). Site-directed mutagenesis using double-stranded plasmid DNA templates. Methods Mol Biol *57*, 31-44.

Brero, A., Leonhardt, H., and Cardoso, M. C. (2006). Replication and translation of epigenetic information. Curr Top Microbiol Immunol *301*, 21-44.

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., Swift, S., and Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell *71*, 515-526.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., and Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527-542.

Buerger, M. J. (1940). The Correction of X-Ray Diffraction Intensities for Lorentz and Polarization Factors. Proc Natl Acad Sci U S A *26*, 637-642.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Moustiakimov, M., Polidori, G., and Siliqi, D. (2004). MAD phasing: choosing the most informative wavelength combination. Acta Crystallogr D Biol Crystallogr *60*, 1683-1686.

Buschdorf, J. P., and Stratling, W. H. (2004). A WW domain binding region in methyl-CpG-binding protein MeCP2: impact on Rett syndrome. J Mol Med *82*, 135-143.

Chadwick, L. H., and Wade, P. A. (2007). MeCP2 in Rett syndrome: transcriptional repressor or chromatin architectural protein? Curr Opin Genet Dev *17*, 121-125.

Chandler, S. P., Guschin, D., Landsberger, N., and Wolffe, A. P. (1999). The methyl-CpG binding transcriptional repressor MeCP2 stably associates with nucleosomal DNA. Biochemistry *38*, 7008-7018.

Chang, Q., Khare, G., Dani, V., Nelson, S., and Jaenisch, R. (2006). The disease progression of Mecp2 mutant mice is affected by the level of BDNF expression. Neuron *49*, 341-348.

Chao, W., Huynh, K. D., Spencer, R. J., Davidow, L. S., and Lee, J. T. (2002). CTCF, a candidate trans-acting factor for X-inactivation choice. Science *295*, 345-347.

Chayen, N. E. (2004). Turning protein crystallisation from an art into a science. Curr Opin Struct Biol *14*, 577-583.

Chayen, N. E. (2005). Methods for separating nucleation and growth in protein crystallisation. Prog Biophys Mol Biol *88*, 329-337.

Chayen, N. E., and Saridakis, E. (2002). Protein crystallization for genomics: towards high-throughput optimization techniques. Acta Crystallogr D Biol Crystallogr *58*, 921-927.

Chen, W. G., Chang, Q., Lin, Y., Meissner, A., West, A. E., Griffith, E. C., Jaenisch, R., and Greenberg, M. E. (2003). Derepression of BDNF transcription involves calcium-dependent phosphorylation of MeCP2. Science *302*, 885-889.

Chen, X., Tordova, M., Gilliland, G. L., Wang, L., Li, Y., Yan, H., and Ji, X. (1998). Crystal structure of apo-cellular retinoic acid-binding protein type II (R111M) suggests a mechanism of ligand entry. J Mol Biol *278*, 641-653.

Colantuoni, C., Jeon, O. H., Hyder, K., Chenchik, A., Khimani, A. H., Narayanan, V., Hoffman, E. P., Kaufmann, W. E., Naidu, S., and Pevsner, J. (2001). Gene expression profiling in postmortem Rett Syndrome brain: differential gene expression and patient classification. Neurobiol Dis *8*, 847-865.

Collaborative (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallographica Section D *50*, 760-763.

Colot, V., and Rossignol, J. L. (1999). Eukaryotic DNA methylation as an evolutionary device. Bioessays *21*, 402-411.

Cooper, D. N., and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. Hum Genet *78*, 151-155.

Couture, J. F., Hauk, G., Thompson, M. J., Blackburn, G. M., and Trievel, R. C. (2006). Catalytic roles for carbon-oxygen hydrogen bonding in SET domain lysine methyltransferases. J Biol Chem *281*, 19280-19287.

Cowtan, K. (1994). 'dm': An automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography *31*, 34-38.

Cowtan, K. D., and Zhang, K. Y. (1999). Density modification for macromolecular phase improvement. Prog Biophys Mol Biol *72*, 245-270.

Cromer, D. T., and Liberman, D. (1970). Relativistic Calculation of Anomalous Scattering Factors for X Rays. The Journal of Chemical Physics *53*, 1891-1898.

Cross, S. H., Meehan, R. R., Nan, X., and Bird, A. (1997). A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. Nat Genet *16*, 256-259.

Daniel, J. M., and Reynolds, A. B. (1999). The catenin p120(ctn) interacts with Kaiso, a novel BTB/POZ domain zinc finger transcription factor. Mol Cell Biol *19*, 3614-3623.

Daniel, J. M., Spring, C. M., Crawford, H. C., Reynolds, A. B., and Baig, A. (2002). The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. Nucleic Acids Res *30*, 2911-2919.

Dauter, Z. (1999). Data-collection strategies. Acta Crystallogr D Biol Crystallogr *55*, 1703-1717.

DeChiara, T. M., Robertson, E. J., and Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. Cell *64*, 849-859.

DeLano, W. L. (2003). The PyMOL molecular graphics system (San Carlos, CA, DeLano Scientific).

Delaval, K., and Feil, R. (2004). Epigenetic regulation of mammalian genomic imprinting. Curr Opin Genet Dev *14*, 188-195.

Deng, V., Matagne, V., Banine, F., Frerking, M., Ohliger, P., Budden, S., Pevsner, J., Dissen, G. A., Sherman, L. S., and Ojeda, S. R. (2007). FXYD1 is an MeCP2 target gene overexpressed in the brains of Rett syndrome patients and Mecp2-null mice. Hum Mol Genet *16*, 640-650.

Derewenda, Z. S. (2004). Rational protein crystallization by mutational surface engineering. Structure *12*, 529-535.

Dong, A., Yoder, J. A., Zhang, X., Zhou, L., Bestor, T. H., and Cheng, X. (2001). Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. Nucleic Acids Res *29*, 439-448.

Dragan, A. I., Liggins, J. R., Crane-Robinson, C., and Privalov, P. L. (2003). The energetics of specific binding of AT-hooks from HMGA1 to target DNA. J Mol Biol *327*, 393-411.

Drenth, J. (2007). Principles of protein X-ray crystallography, 3rd edn (New York, Springer Science+Business Media, LLC).

Drew, H. R., Wing, R. M., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1981). Structure of a B-DNA dodecamer: conformation and dynamics. Proc Natl Acad Sci U S A *78*, 2179-2183.

Drewell, R. A., Goddard, C. J., Thomas, J. O., and Surani, M. A. (2002). Methylation-dependent silencing at the H19 imprinting control region by MeCP2. Nucleic Acids Res *30*, 1139-1144.

Durbin, S. D., and Feher, G. (1996). Protein crystallization. Annu Rev Phys Chem *47*, 171-204.

El Hassan, M. A., and Calladine, C. R. (1996). Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. J Mol Biol *259*, 95-103.

El Hassan, M. A., and Calladine, C. R. (1998). Two distinct modes of protein-induced bending in DNA. J Mol Biol *282*, 331-343.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr *60*, 2126-2132.

Engh, R. A., and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallographica Section A *47*, 392-400.

Evans, P. (2006). Scaling and assessment of data quality. Acta Crystallogr D Biol Crystallogr *62*, 72-82.

Feil, R., and Berger, F. (2007). Convergent evolution of genomic imprinting in plants and mammals. Trends Genet *23*, 192-199.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature *421*, 448-453.

Feng, Q., and Zhang, Y. (2001). The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. Genes Dev *15*, 827-832.

Fraga, M. F., Ballestar, E., Montoya, G., Taysavang, P., Wade, P. A., and Esteller, M. (2003). The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. Nucleic Acids Res *31*, 1765-1774.

Fratini, A. V., Kopka, M. L., Drew, H. R., and Dickerson, R. E. (1982). Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTBrCGCG. J Biol Chem *257*, 14686-14707.

Free, A., Wakefield, R. I., Smith, B. O., Dryden, D. T., Barlow, P. N., and Bird, A. P. (2001). DNA recognition by the methyl-CpG binding domain of MeCP2. J Biol Chem *276*, 3353-3360.

Fujita, N., Takebayashi, S., Okumura, K., Kudo, S., Chiba, T., Saya, H., and Nakao, M. (1999). Methylation-mediated transcriptional silencing in euchromatin by methyl-CpG binding protein MBD1 isoforms. Mol Cell Biol *19*, 6415-6426.

Fujita, N., Watanabe, S., Ichimura, T., Ohkuma, Y., Chiba, T., Saya, H., and Nakao, M. (2003a). MCAF mediates MBD1-dependent transcriptional repression. Mol Cell Biol *23*, 2834-2843.

Fujita, N., Watanabe, S., Ichimura, T., Tsuruzoe, S., Shinkai, Y., Tachibana, M., Chiba, T., and Nakao, M. (2003b). Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. J Biol Chem *278*, 24132-24138.

Fuks, F., Hurd, P. J., Deplus, R., and Kouzarides, T. (2003a). The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. Nucleic Acids Res *31*, 2305-2312.

Fuks, F., Hurd, P. J., Wolf, D., Nan, X., Bird, A. P., and Kouzarides, T. (2003b). The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. J Biol Chem *278*, 4035-4040.

Garman, E., and Murray, J. W. (2003). Heavy-atom derivatization. Acta Crystallogr D Biol Crystallogr *59*, 1903-1913.

Garman, E. F., and Owen, R. L. (2006). Cryocooling and radiation damage in macromolecular crystallography. Acta Crystallogr D Biol Crystallogr *62*, 32-47.

Georgel, P. T., Horowitz-Scherer, R. A., Adkins, N., Woodcock, C. L., Wade, P. A., and Hansen, J. C. (2003). Chromatin compaction by human MeCP2. Assembly of novel secondary chromatin structures in the absence of DNA methylation. J Biol Chem *278*, 32181-32188.

Goll, M. G., Kirpekar, F., Maggert, K. A., Yoder, J. A., Hsieh, C. L., Zhang, X., Golic, K. G., Jacobsen, S. E., and Bestor, T. H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science *311*, 395-398.

Golovin, A., Oldfield, T. J., Tate, J. G., Velankar, S., Barton, G. J., Boutselakis, H., Dimitropoulos, D., Fillon, J., Hussain, A., Ionides, J. M.*, et al.* (2004). E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res *32*, D211-216.

Gouet, P., Courcelle, E., Stuart, D. I., and Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics *15*, 305-308.

Gowher, H., and Jeltsch, A. (2001). Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG [correction of non-CpA] sites. J Mol Biol *309*, 1201-1208.

Guy, J., Gan, J., Selfridge, J., Cobb, S., and Bird, A. (2007). Reversal of neurological defects in a mouse model of Rett syndrome. Science *315*, 1143-1147.

Hagberg, B. (1985). Rett's syndrome: prevalence and impact on progressive severe mental retardation in girls. Acta Paediatr Scand *74*, 405-408.

Hagberg, B., Aicardi, J., Dias, K., and Ramos, O. (1983). A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases. Ann Neurol *14*, 471-479.

Harikrishnan, K. N., Chow, M. Z., Baker, E. K., Pal, S., Bassal, S., Brasacchio, D., Wang, L., Craig, J. M., Jones, P. L., Sif, S., and El-Osta, A. (2005). Brahma links the SWI/SNF chromatin-remodeling complex with MeCP2-dependent transcriptional silencing. Nat Genet *37*, 254-264.

Hauptman, H. (1997a). Phasing methods for protein crystallography. Curr Opin Struct Biol *7*, 672-680.

Hauptman, H. A. (1997b). Shake-and-bake: an algorithm for automatic solution ab initio of crystal structures. Methods Enzymol *277*, 3-13.

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol *18*, 6538-6547.

Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V. A., and Bird, A. (2001). Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. Genes Dev *15*, 710-723.

Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature *401*, 301-304.

Hendrich, B., and Tweedie, S. (2003). The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet *19*, 269-277.

Hendrickson, W. A., Horton, J. R., and LeMaster, D. M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. Embo J *9*, 1665-1672.

Hendrickson, W. A., Smith, J. L., and Sheriff, S. (1985). Direct phase determination based on anomalous scattering. Methods Enzymol *115*, 41-55.

Hermann, A., Gowher, H., and Jeltsch, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases. Cell Mol Life Sci *61*, 2571-2587.

Hermann, A., Schmitt, S., and Jeltsch, A. (2003). The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. J Biol Chem *278*, 31717-31721.

Ho, K. L., McNae, I. W., Schmiedeberg, L., Klose, R. J., Bird, A. P., and Walkinshaw, M. D. (2008). MeCP2 binding to DNA depends upon hydration at methyl-CpG. Mol Cell *29*, 525-531.

Horike, S., Cai, S., Miyano, M., Cheng, J. F., and Kohwi-Shigematsu, T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet *37*, 31-40.

Hutchinson, E. G., and Thornton, J. M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. Protein Sci *5*, 212-220.

Huth, J. R., Bewley, C. A., Nissen, M. S., Evans, J. N., Reeves, R., Gronenborn, A. M., and Clore, G. M. (1997). The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. Nat Struct Biol *4*, 657-665.

Ichimura, T., Watanabe, S., Sakamoto, Y., Aoto, T., Fujita, N., and Nakao, M. (2005). Transcriptional repression and heterochromatin formation by MBD1 and MCAF/AM family proteins. J Biol Chem *280*, 13928-13935.

Itoh, M., Ide, S., Takashima, S., Kudo, S., Nomura, Y., Segawa, M., Kubota, T., Mori, H., Tanaka, S., Horie, H.*, et al.* (2007). Methyl CpG-binding protein 2 (a mutation of which causes Rett syndrome) directly regulates insulin-like growth factor binding protein 3 in mouse and human brains. J Neuropathol Exp Neurol *66*, 117-123.

Jeffery, L., and Nakielny, S. (2004). Components of the DNA methylation system of chromatin control are RNA-binding proteins. J Biol Chem *279*, 49479-49487.

Jones, P. L., Veenstra, G. J., Wade, P. A., Vermaak, D., Kass, S. U., Landsberger, N., Strouboulis, J., and Wolffe, A. P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet *19*, 187-191.

Jorgensen, H. F., Ben-Porath, I., and Bird, A. P. (2004). Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. Mol Cell Biol *24*, 3387-3395.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.

Kaludov, N. K., and Wolffe, A. P. (2000). MeCP2 driven transcriptional repression in vitro: selectivity for methylated DNA, action at a distance and contacts with the basal transcription machinery. Nucleic Acids Res *28*, 1921-1928.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for *de novo* DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900-903.

Kim, S. W., Fang, X., Ji, H., Paulson, A. F., Daniel, J. M., Ciesiolka, M., van Roy, F., and McCrea, P. D. (2002). Isolation and characterization of XKaiso, a transcriptional repressor that associates with the catenin Xp120(ctn) in Xenopus laevis. J Biol Chem *277*, 8202-8208.

Kimura, H., and Shiota, K. (2003). Methyl-CpG-binding protein, MeCP2, is a target molecule for maintenance DNA methyltransferase, Dnmt1. J Biol Chem *278*, 4806-4812.

Kishi, N., and Macklis, J. D. (2004). MECP2 is progressively expressed in post-migratory neurons and is involved in neuronal maturation rather than cell fate decisions. Mol Cell Neurosci *27*, 306-321.

Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K., and Jones, T. A. (1994). Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure *2*, 1241-1258.

Klimasauskas, S., Kumar, S., Roberts, R. J., and Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. Cell *76*, 357-369.

Klose, R. J., and Bird, A. P. (2004). MeCP2 behaves as an elongated monomer that does not stably associate with the Sin3a chromatin remodeling complex. J Biol Chem *279*, 46490-46496.

Klose, R. J., and Bird, A. P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci *31*, 89-97.

Klose, R. J., Sarraf, S. A., Schmiedeberg, L., McDermott, S. M., Stancheva, I., and Bird, A. P. (2005). DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. Mol Cell *19*, 667-678.

Kochanek, S., Renz, D., and Doerfler, W. (1995). Transcriptional silencing of human Alu sequences and inhibition of protein binding in the box B regulatory elements by 5'-CG-3' methylation. FEBS Lett *360*, 115-120.

Kokura, K., Kaul, S. C., Wadhwa, R., Nomura, T., Khan, M. M., Shinagawa, T., Yasukawa, T., Colmenares, C., and Ishii, S. (2001). The Ski protein family is required for MeCP2-mediated transcriptional repression. J Biol Chem *276*, 34115-34121.

Kondo, E., Gu, Z., Horii, A., and Fukushige, S. (2005). The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. Mol Cell Biol *25*, 4388-4396.

Kouzarides, T. (2002). Histone methylation in transcriptional control. Curr Opin Genet Dev *12*, 198-209.

Krawczak, M., Ball, E. V., and Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet *63*, 474-488.

Kriaucionis, S., and Bird, A. (2003). DNA methylation and Rett syndrome. Hum Mol Genet *12 Spec No 2*, R221-227.

Kriaucionis, S., and Bird, A. (2004). The major form of MeCP2 has a novel N-terminus generated by alternative splicing. Nucleic Acids Res *32*, 1818-1823.

Krithivas, A., Fujimuro, M., Weidner, M., Young, D. B., and Hayward, S. D. (2002). Protein interactions targeting the latency-associated nuclear antigen of Kaposi's sarcoma-associated herpesvirus to cell chromosomes. J Virol *76*, 11596-11604.

Kunkel, T. A. (1985). Rapid and efficient site-specific mutagenesis without phenotypic selection. Proc Natl Acad Sci U S A *82*, 488-492.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography *26*, 283-291.

Lauster, R., Trautner, T. A., and Noyer-Weidner, M. (1989). Cytosine-specific type II DNA methyltransferases. A conserved enzyme core with variable target-recognizing domains. J Mol Biol *206*, 305-312.

Le Guezennec, X., Vermeulen, M., Brinkman, A. B., Hoeijmakers, W. A., Cohen, A., Lasonder, E., and Stunnenberg, H. G. (2006). MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. Mol Cell Biol *26*, 843-851.

Leahy, D. J., Erickson, H. P., Aukhil, I., Joshi, P., and Hendrickson, W. A. (1994). Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine. Proteins *19*, 48-54.

Lee, J. H., Voo, K. S., and Skalnik, D. G. (2001). Identification and characterization of the DNA binding domain of CpG-binding protein. J Biol Chem *276*, 44669-44676.

Lee, J. T., and Lu, N. (1999). Targeted mutagenesis of Tsix leads to nonrandom X inactivation. Cell *99*, 47-57.

Leslie, A. (1999). Integration of macromolecular diffraction data. Acta Crystallographica Section D *55*, 1696-1702.

Leslie, A. G. (2006). The integration of macromolecular diffraction data. Acta Crystallogr D Biol Crystallogr *62*, 48-57.

Leslie, A. G. W. (1992). Joint CCP4-ESF-EAMCB Newsletter on Protein Crystallography *26*, 22-33.

Lewis, J. D., Meehan, R. R., Henzel, W. J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell *69*, 905-914.

Liang, G., Chan, M. F., Tomigahara, Y., Tsai, Y. C., Gonzales, F. A., Li, E., Laird, P. W., and Jones, P. A. (2002). Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Mol Cell Biol *22*, 480-491.

Lu, X. J., and Olson, W. K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res *31*, 5108-5121.

Luft, J. R., Wolfley, J., Jurisica, I., Glasgow, J., Fortier, S., and DeTitta, G. T. (2001). Macromolecular crystallization in a high throughput laboratory--the search phase. Journal of Crystal Growth *232*, 591-595.

Lunyak, V. V., Burgess, R., Prefontaine, G. G., Nelson, C., Sze, S. H., Chenoweth, J., Schwartz, P., Pevzner, P. A., Glass, C., Mandel, G., and Rosenfeld, M. G. (2002). Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. Science *298*, 1747-1752.

Luscombe, N. M., Laskowski, R. A., and Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res *29*, 2860-2874.

Lyon, M. F. (1961). Gene action in the X-chromosome of the mouse (Mus musculus L.). Nature *190*, 372-373.

Mack, D. R., Chiu, T. K., and Dickerson, R. E. (2001). Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. J Mol Biol *312*, 1037-1049.

Magdinier, F., and Wolffe, A. P. (2001). Selective association of the methyl-CpG binding protein MBD2 with the silent p14/p16 locus in human neoplasia. Proc Natl Acad Sci U S A *98*, 4990-4995.

Marahrens, Y., Panning, B., Dausman, J., Strauss, W., and Jaenisch, R. (1997). Xist-deficient mice are defective in dosage compensation but not spermatogenesis. Genes Dev *11*, 156-166.

Margot, J. B., Aguirre-Arteta, A. M., Di Giacco, B. V., Pradhan, S., Roberts, R. J., Cardoso, M. C., and Leonhardt, H. (2000). Structure and function of the mouse DNA methyltransferase gene: Dnmt1 shows a tripartite structure. J Mol Biol *297*, 293-300.

Mari, F., Azimonti, S., Bertani, I., Bolognese, F., Colombo, E., Caselli, R., Scala, E., Longo, I., Grosso, S., Pescucci, C.*, et al.* (2005). CDKL5 belongs to the same molecular pathway of MeCP2 and it is responsible for the early-onset seizure variant of Rett syndrome. Hum Mol Genet *14*, 1935-1946.

Martinowich, K., Hattori, D., Wu, H., Fouse, S., He, F., Hu, Y., Fan, G., and Sun, Y. E. (2003). DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. Science *302*, 890-893.

Matthews, B. W. (1968). Solvent content of protein crystals. J Mol Biol *33*, 491-497.

Mayer-Jung, C., Moras, D., and Timsit, Y. (1998). Hydration and recognition of methylated CpG steps in DNA. Embo J *17*, 2709-2718.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007). Phaser crystallographic software. Journal of Applied Crystallography *40*, 658-674.

McDonald, I. K., and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. J Mol Biol *238*, 777-793.

McGrath, J., and Solter, D. (1984). Completion of mouse embryogenesis requires both the maternal and paternal genomes. Cell *37*, 179-183.

McPherson, A. (1999). Crystallization of biological macromolecules (New York, Cold Spring Harbor Laboratory Press).

McPherson, A. (2004). Introduction to protein crystallization. Methods *34*, 254-265.

Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L., and Bird, A. P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell *58*, 499-507.

Millar, C. B., Guy, J., Sansom, O. J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P. D., Bishop, S. M., Clarke, A. R., and Bird, A. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science *297*, 403-405.

Miranda, T. B., and Jones, P. A. (2007). DNA methylation: The nuts and bolts of repression. J Cell Physiol *213*, 384-390.

Mnatzakanian, G. N., Lohi, H., Munteanu, I., Alfred, S. E., Yamada, T., MacLeod, P. J., Jones, J. R., Scherer, S. W., Schanen, N. C., Friez, M. J.*, et al.* (2004). A previously unidentified MECP2 open reading frame defines a new protein isoform relevant to Rett syndrome. Nat Genet *36*, 339-341.

Mochalkin, I., Cheng, B., Klezovitch, O., Scanu, A. M., and Tulinsky, A. (1999). Recombinant kringle IV-10 modules of human apolipoprotein(a): structure, ligand binding modes, and biological relevance. Biochemistry *38*, 1990-1998.

Morris, R. J., Perrakis, A., and Lamzin, V. S. (2003). ARP/wARP and automatic interpretation of protein electron density maps. Methods Enzymol *374*, 229-244.

Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A., and Lamzin, V. S. (2004). Breaking good resolutions with ARP/wARP. J Synchrotron Radiat *11*, 56-59.

Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr *53*, 240-255.

Nan, X., Campoy, F. J., and Bird, A. (1997). MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell *88*, 471-481.

Nan, X., Meehan, R. R., and Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Res *21*, 4886-4892.

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Nan, X., Tate, P., Li, E., and Bird, A. (1996). DNA methylation specifies chromosomal localization of MeCP2. Mol Cell Biol *16*, 414-421.

Neddermann, P., Gallinari, P., Lettieri, T., Schmid, D., Truong, O., Hsuan, J. J., Wiebauer, K., and Jiricny, J. (1996). Cloning and expression of human G/T mismatch-specific thymine-DNA glycosylase. J Biol Chem *271*, 12767-12774.

Nelson, H. C., Finch, J. T., Luisi, B. F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature *330*, 221-226.

Ng, H. H., Jeppesen, P., and Bird, A. (2000). Active repression of methylated genes by the chromosomal protein MBD1. Mol Cell Biol *20*, 1394-1406.

Ng, H. H., Zhang, Y., Hendrich, B., Johnson, C. A., Turner, B. M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., and Bird, A. (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. Nat Genet *23*, 58-61.

Nilsen, H., Haushalter, K. A., Robins, P., Barnes, D. E., Verdine, G. L., and Lindahl, T. (2001). Excision of deaminated cytosine from the vertebrate genome: role of the SMUG1 uracil-DNA glycosylase. Embo J *20*, 4278-4286.

Norris, D. P., Patel, D., Kay, G. F., Penny, G. D., Brockdorff, N., Sheardown, S. A., and Rastan, S. (1994). Evidence that random and imprinted Xist expression is controlled by preemptive methylation. Cell *77*, 41-51.

Nuber, U. A., Kriaucionis, S., Roloff, T. C., Guy, J., Selfridge, J., Steinhoff, C., Schulz, R., Lipkowitz, B., Ropers, H. H., Holmes, M. C., and Bird, A. (2005). Up-regulation of glucocorticoid-regulated genes in a mouse model of Rett syndrome. Hum Mol Genet *14*, 2247-2256.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G., and Notredame, C. (2004). 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J Mol Biol *340*, 385-395.

Ohhata, T., Hoki, Y., Sasaki, H., and Sado, T. (2008). Crucial role of antisense transcription across the Xist promoter in Tsix-mediated Xist chromatin modification. Development *135*, 227-235.

Ohki, I., Shimotake, N., Fujita, N., Jee, J., Ikegami, T., Nakao, M., and Shirakawa, M. (2001). Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. Cell *105*, 487-497.

Ohki, I., Shimotake, N., Fujita, N., Nakao, M., and Shirakawa, M. (1999). Solution structure of the methyl-CpG-binding domain of the methylation-dependent transcriptional repressor MBD1. EMBO J *18*, 6653-6661.

Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. Cell *99*, 247-257.

Okano, M., Xie, S., and Li, E. (1998a). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat Genet *19*, 219-220.

Okano, M., Xie, S., and Li, E. (1998b). Dnmt2 is not required for *de novo* and maintenance methylation of viral DNA in embryonic stem cells. Nucleic Acids Res *26*, 2536-2540.

Ooi, S. K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S. P., Allis, C. D.*, et al.* (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to *de novo* methylation of DNA. Nature *448*, 714-717.

Otwinowski, Z., Borek, D., Majewski, W., and Minor, W. (2003). Multiparametric scaling of diffraction intensities. Acta Crystallogr A *59*, 228-234.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol *276*, 307-326.

Painter, J., and Merritt, E. A. (2006). TLSMD web server for the generation of multi-group TLS models. Journal of Applied Crystallography *39*, 109-111.

Patterson, A. L. (1934). A Fourier Series Method for the Determination of the Components of Interatomic Distances in Crystals. Physical Review *46*, 372.

Peddada, S., Yasui, D. H., and LaSalle, J. M. (2006). Inhibitors of differentiation (ID1, ID2, ID3 and ID4) genes are neuronal targets of MeCP2 that are elevated in Rett syndrome. Hum Mol Genet *15*, 2003-2014.

Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. Nature *379*, 131-137.

Pflugrath, J. W. (1999). The finer things in X-ray diffraction data collection. Acta Crystallogr D Biol Crystallogr *55*, 1718-1725.

Posfai, J., Bhagwat, A. S., Posfai, G., and Roberts, R. J. (1989). Predictive motifs derived from cytosine methyltransferases. Nucleic Acids Res *17*, 2421-2435.

Potterton, E., McNicholas, S., Krissinel, E., Cowtan, K., and Noble, M. (2002). The CCP4 molecular-graphics project. Acta Crystallogr D Biol Crystallogr *58*, 1955-1957.

Prive, G. G., Heinemann, U., Chandrasegaran, S., Kan, L. S., Kopka, M. L., and Dickerson, R. E. (1987). Helix geometry, hydration, and G.A mismatch in a B-DNA decamer. Science *238*, 498-504.

Prokhortchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., and Prokhortchouk, E. (2001). The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes Dev *15*, 1613-1618.

Prokhortchouk, A., Sansom, O., Selfridge, J., Caballero, I. M., Salozhin, S., Aithozhina, D., Cerchietti, L., Meng, F. G., Augenlicht, L. H., Mariadason, J. M.*, et al.* (2006). Kaiso-deficient mice show resistance to intestinal cancer. Mol Cell Biol *26*, 199-208.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. J Mol Biol *7*, 95-99.

Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc Natl Acad Sci U S A *97*, 5237-5242.

Reese, K. J., Lin, S., Verona, R. I., Schultz, R. M., and Bartolomei, M. S. (2007). Maintenance of paternal methylation and repression of the imprinted H19 gene requires MBD3. PLoS Genet *3*, e137.

Reinisch, K. M., Chen, L., Verdine, G. L., and Lipscomb, W. N. (1995). The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. Cell *82*, 143-153.

Rett, A. (1966). On a unusual brain atrophy syndrome in hyperammonemia in childhood. Wien Med Wochenschr *116*, 723-726.

Rhee, I., Bachman, K. E., Park, B. H., Jair, K. W., Yen, R. W., Schuebel, K. E., Cui, H., Feinberg, A. P., Lengauer, C., Kinzler, K. W.*, et al.* (2002). DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. Nature *416*, 552-556.

Rhee, I., Jair, K. W., Yen, R. W., Lengauer, C., Herman, J. G., Kinzler, K. W., Vogelstein, B., Baylin, S. B., and Schuebel, K. E. (2000). CpG methylation is maintained in human cancer cells lacking DNMT1. Nature *404*, 1003-1007.

Robertson, K. D., and Wolffe, A. P. (2000). DNA methylation in health and disease. Nat Rev Genet *1*, 11-19.

Rountree, M. R., Bachman, K. E., and Baylin, S. B. (2000). DNMT1 binds HDAC2 and a new co-repressor, DMAP1, to form a complex at replication foci. Nat Genet *25*, 269-277.

Ruzov, A., Dunican, D. S., Prokhortchouk, A., Pennings, S., Stancheva, I., Prokhortchouk, E., and Meehan, R. R. (2004). Kaiso is a genome-wide repressor of transcription that is essential for amphibian development. Development *131*, 6185-6194.

Sado, T., Hoki, Y., and Sasaki, H. (2005). Tsix silences Xist through modification of chromatin structure. Dev Cell *9*, 159-165.

Sansom, O. J., Berger, J., Bishop, S. M., Hendrich, B., Bird, A., and Clarke, A. R. (2003). Deficiency of Mbd2 suppresses intestinal tumorigenesis. Nat Genet *34*, 145-147.

Sansom, O. J., Maddison, K., and Clarke, A. R. (2007). Mechanisms of disease: methyl-binding domain proteins as potential therapeutic targets in cancer. Nat Clin Pract Oncol *4*, 305-315.

Sarraf, S. A., and Stancheva, I. (2004). Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. Mol Cell *15*, 595-605.

Schneider, T. R., and Sheldrick, G. M. (2002). Substructure solution with SHELXD. Acta Crystallogr D Biol Crystallogr *58*, 1772-1779.

Selker, E. U. (1997). Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion? Trends Genet *13*, 296-301.

Sharff, A. J., Koronakis, E., Luisi, B., and Koronakis, V. (2000). Oxidation of selenomethionine: some MADness in the method! Acta Crystallogr D Biol Crystallogr *56*, 785-788.

Sheldrick, G. M. (2008). A short history of SHELX. Acta Crystallogr A *64*, 112-122.

Stancheva, I., Collins, A. L., Van den Veyver, I. B., Zoghbi, H., and Meehan, R. R. (2003). A mutant form of MeCP2 protein associated with human Rett syndrome cannot be displaced from methylated DNA by notch in Xenopus embryos. Mol Cell *12*, 425-435.

Stavropoulos, N., Lu, N., and Lee, J. T. (2001). A functional role for Tsix transcription in blocking Xist RNA accumulation but not in X-chromosome choice. Proc Natl Acad Sci U S A *98*, 10232-10237.

Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V., and Feigon, J. (2004). DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. Proc Natl Acad Sci U S A *101*, 1177-1182.

Stoscheck, C. M. (1990). Quantitation of protein. Methods Enzymol *182*, 50-68.

Suetake, I., Miyazaki, J., Murakami, C., Takeshima, H., and Tajima, S. (2003). Distinct enzymatic properties of recombinant mouse DNA methyltransferases Dnmt3a and Dnmt3b. J Biochem (Tokyo) *133*, 737-744.

Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004). DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. J Biol Chem *279*, 27816-27823.

Sugimoto, M., Esaki, N., Tanaka, H., and Soda, K. (1989). A simple and efficient method for the oligonucleotide-directed mutagenesis using plasmid DNA template and phosphorothioate-modified nucleotide. Anal Biochem *179*, 309-311.

Surani, M. A., Barton, S. C., and Norris, M. L. (1984). Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. Nature *308*, 548-550.

Suzuki, H., Watkins, D. N., Jair, K. W., Schuebel, K. E., Markowitz, S. D., Chen, W. D., Pretlow, T. P., Yang, B., Akiyama, Y., Van Engeland, M*., et al.* (2004). Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. Nat Genet *36*, 417-422.

Suzuki, M., Yamada, T., Kihara-Negishi, F., Sakurai, T., and Oikawa, T. (2003). Direct association between PU.1 and MeCP2 that recruits mSin3A-HDAC complex for PU.1-mediated transcriptional repression. Oncogene *22*, 8688-8698.

Tariq, M., and Paszkowski, J. (2004). DNA and histone methylation in plants. Trends Genet *20*, 244-251.

Tatematsu, K. I., Yamazaki, T., and Ishikawa, F. (2000). MBD2-MBD3 complex binds to hemi-methylated DNA and forms a complex containing DNMT1 at the replication foci in late S phase. Genes Cells *5*, 677-688.

Taylor, G. (2003). The phase problem. Acta Crystallogr D Biol Crystallogr *59*, 1881-1890.

Taylor, J. W., Ott, J., and Eckstein, F. (1985). The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA. Nucleic Acids Res *13*, 8765-8785.

Terwilliger, T. (2004). SOLVE and RESOLVE: automated structure solution, density modification and model building. J Synchrotron Radiat *11*, 49-52.

Terwilliger, T. C. (2003a). Automated main-chain model building by template matching and iterative fragment extension. Acta Crystallogr D Biol Crystallogr *59*, 38-44.

Terwilliger, T. C. (2003b). Automated side-chain model building and sequence assignment by template matching. Acta Crystallogr D Biol Crystallogr *59*, 45-49.

Terwilliger, T. C. (2003c). SOLVE and RESOLVE: automated structure solution and density modification. Methods Enzymol *374*, 22-37.

Terwilliger, T. C., and Eisenberg, D. (1983). Unbiased three-dimensional refinement of heavy-atom parameters by correlation of origin-removed Patterson functions. Acta Crystallographica Section A *39*, 813-817.

Towbin, H., Staehelin, T., and Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proc Natl Acad Sci U S A *76*, 4350-4354.

Traynor, J., Agarwal, P., Lazzeroni, L., and Francke, U. (2002). Gene expression patterns vary in clonal cell cultures from Rett syndrome females with eight different MECP2 mutations. BMC Med Genet *3*, 12.

Tronrud, D. E. (2004). Introduction to macromolecular refinement. Acta Crystallogr D Biol Crystallogr *60*, 2156-2168.

Tudor, M., Akbarian, S., Chen, R. Z., and Jaenisch, R. (2002). Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain. Proc Natl Acad Sci U S A *99*, 15536-15541.

Uson, I., Schmidt, B., von Bulow, R., Grimme, S., von Figura, K., Dauter, M., Rajashankar, K. R., Dauter, Z., and Sheldrick, G. M. (2003). Locating the anomalous scatterer substructures in halide and sulfur phasing. Acta Crystallogr D Biol Crystallogr *59*, 57-66.

Vaguine, A. A., Richelle, J., and Wodak, S. J. (1999). SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Crystallographica Section D *55*, 191-205.

Valinluck, V., Liu, P., Kang, J. I., Jr., Burdzy, A., and Sowers, L. C. (2005). 5-halogenated pyrimidine lesions within a CpG sequence context mimic 5-methylcytosine by enhancing the binding of the methyl-CpG-binding domain of methyl-CpG-binding protein 2 (MeCP2). Nucleic Acids Res *33*, 3057-3064.

Valinluck, V., Tsai, H. H., Rogstad, D. K., Burdzy, A., Bird, A., and Sowers, L. C. (2004). Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). Nucleic Acids Res *32*, 4100-4108.

Vandeyar, M. A., Weiner, M. P., Hutton, C. J., and Batt, C. A. (1988). A simple and rapid method for the selection of oligodeoxynucleotide-directed mutants. Gene *65*, 129-133.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A*., et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Verona, R. I., Mann, M. R., and Bartolomei, M. S. (2003). Genomic imprinting: intricacies of epigenetic regulation in clusters. Annu Rev Cell Dev Biol *19*, 237-259.

Vonrhein, C., Blanc, E., Roversi, P., and Bricogne, G. (2006). Automated Structure Solution With autoSHARP. Methods Mol Biol *364*, 215-230.

Wade, P. A., Gegonne, A., Jones, P. L., Ballestar, E., Aubry, F., and Wolffe, A. P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. Nat Genet *23*, 62-66.

Wakefield, R. I., Smith, B. O., Nan, X., Free, A., Soteriou, A., Uhrin, D., Bird, A. P., and Barlow, P. N. (1999). The solution structure of the domain from MeCP2 that binds to methylated DNA. J Mol Biol *291*, 1055-1065.

Wan, W. Y., and Milner-White, E. J. (1999a). A natural grouping of motifs with an aspartate or asparagine residue forming two hydrogen bonds to residues ahead in sequence: their occurrence at alpha-helical N termini and in other situations. J Mol Biol *286*, 1633-1649.

Wan, W. Y., and Milner-White, E. J. (1999b). A recurring two-hydrogen-bond motif incorporating a serine or threonine residue is found both at alpha-helical N termini and in other situations. J Mol Biol *286*, 1651-1662.

Wang, B. C. (1985). Resolution of phase ambiguity in macromolecular crystallography. In Diffraction methods for biological macromolecules, H. W. Wyckokoff, Hirs, C.H.W., Timasheff, S.N., ed. (Orlando, Academic Press), pp. 90-113.

Watanabe, S., Ichimura, T., Fujita, N., Tsuruzoe, S., Ohki, I., Shirakawa, M., Kawasuji, M., and Nakao, M. (2003). Methylated DNA-binding domain 1 and methylpurine-DNA glycosylase link transcriptional repression and DNA repair in chromatin. Proc Natl Acad Sci U S A *100*, 12859-12864.

Watson, J. D., and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Weeks, C. M., and Miller, R. (1999). Optimizing Shake-and-Bake for proteins. Acta Crystallogr D Biol Crystallogr *55*, 492-500.

Wiebauer, K., and Jiricny, J. (1989). In vitro correction of G.T mispairs to G.C pairs in nuclear extracts from human cells. Nature *339*, 234-236.

Wilkinson, C. R., Bartlett, R., Nurse, P., and Bird, A. P. (1995). The fission yeast gene pmt1+ encodes a DNA methyltransferase homologue. Nucleic Acids Res *23*, 203-210.

Wilson, G. G., and Murray, N. E. (1991). Restriction and modification systems. Annu Rev Genet *25*, 585-627.

Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1980). Crystal structure analysis of a complete turn of B-DNA. Nature *287*, 755-758.

Wong, E., Yang, K., Kuraguchi, M., Werling, U., Avdievich, E., Fan, K., Fazzari, M., Jin, B., Brown, A. M., Lipkin, M., and Edelmann, W. (2002). Mbd4 inactivation increases Cright-arrowT transition mutations and promotes gastrointestinal tumor formation. Proc Natl Acad Sci U S A *99*, 14937-14942.

Wutz, A., and Jaenisch, R. (2000). A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. Mol Cell *5*, 695-705.

Xie, S., Wang, Z., Okano, M., Nogami, M., Li, Y., He, W. W., Okumura, K., and Li, E. (1999). Cloning, expression and chromosome locations of the human DNMT3 gene family. Gene *236*, 87-95.

Yoder, J. A., and Bestor, T. H. (1998). A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. Hum Mol Genet *7*, 279-284.

Yoder, J. A., Soman, N. S., Verdine, G. L., and Bestor, T. H. (1997a). DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe. J Mol Biol *270*, 385-395.

Yoder, J. A., Walsh, C. P., and Bestor, T. H. (1997b). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet *13*, 335-340.

Yoon, H. G., Chan, D. W., Reynolds, A. B., Qin, J., and Wong, J. (2003). N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. Mol Cell *12*, 723-734.

Young, J. I., Hong, E. P., Castle, J. C., Crespo-Barreto, J., Bowman, A. B., Rose, M. F., Kang, D., Richman, R., Johnson, J. M., Berget, S., and Zoghbi, H. Y. (2005). Regulation of RNA splicing by the methylation-dependent transcriptional repressor methyl-CpG binding protein 2. Proc Natl Acad Sci U S A *102*, 17551-17558.

Yusufzai, T. M., and Wolffe, A. P. (2000). Functional consequences of Rett syndrome mutations on human MeCP2. Nucleic Acids Res *28*, 4172-4179.

Zhang, K. Y. J., and Main, P. (1990). The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. Acta Crystallographica Section A *46*, 377-381.

Zhang, Y., Ng, H. H., Erdjument-Bromage, H., Tempst, P., Bird, A., and Reinberg, D. (1999). Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes Dev *13*, 1924-1935.

Zhou, Z., Hong, E. J., Cohen, S., Zhao, W. N., Ho, H. Y., Schmidt, L., Chen, W. G., Lin, Y., Savner, E., Griffith, E. C.*, et al.* (2006). Brain-specific phosphorylation of MeCP2 regulates activity-dependent Bdnf transcription, dendritic growth, and spine maturation. Neuron *52*, 255-269.

Zimmermann, C., Guhl, E., and Graessmann, A. (1997). Mouse DNA methyltransferase (MTase) deletion mutants that retain the catalytic domain display neither *de novo* nor maintenance methylation activity *in vivo*. Biol Chem *378*, 393-405.

PUBLICATIONS

1. **Ho, K.L.**, McNae, I.W., Schmiedeberg, L., Klose R.J., Bird, A.B., and Walkinshaw, W.D. (2008). MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Molecular Cell*. ***29***: 525-531.

2. Tan, W.S., McNae, I.W., **Ho, K.L.**, and Walkinshaw, M.D. (2007). Crystallization and X-ray Analysis of the T=4 particle Hepatitis B capsid proten with an N-terminal Extension. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2007 Aug 1;63(Pt 8):642-7

3. **Ho, K.L.**, Yusoff, K., Seow, H.F., and Tan, W.S. (2003). Selection of high affinity ligands to Hepatitis B core particles from a phage-displayed cyclic peptide library. *Journal of Medical Virology*. **69**: 27-32.